1. Markov chains

Section 1. What is a Markov chain? How to simulate one.
Section 2. The Markov property.
Section 3. How matrix multiplication gets into the picture.
Section 4. Statement of the Basic Limit Theorem about convergence to stationarity. A motivating example shows how complicated random objects can be generated using Markov chains.
Section 5. Stationary distributions, with examples. An exercise introduces the idea of probability flux.
Section 6. Other concepts from the Basic Limit Theorem: irreducibility, periodicity, and recurrence. An interesting classical example: recurrence or transience of random walks.
Section 7. Introduces the idea of coupling.
Section 8. Uses coupling to prove the Basic Limit Theorem.
Section 9. A Strong Law of Large Numbers for Markov chains.
Section 10. Markov chains in general state spaces.

Markov chains are a relatively simple but very interesting and useful class of random processes. A Markov chain describes a system whose state changes over time. The changes are not completely predictable, but rather are governed by probability distributions. These probability distributions incorporate a simple sort of dependence structure, where the conditional distribution of future states of the system, given some information about past states, depends only on the most recent piece of information. That is, what matters in predicting the future of the system is its present state, and not the path by which the system got to its present state. Markov chains illustrate many of the important ideas of stochastic processes in an elementary setting. This classical subject is still very much alive, with important developments in both theory and applications coming at an accelerating pace in recent decades.

1.1 Specifying and simulating a Markov chain

What is a Markov chain*? One answer is to say that it is a sequence $\{X_0, X_1, X_2, \ldots\}$ of random variables that has the "Markov property"; we will discuss this in the next section. For now, to get a feeling for what a Markov chain is, let's think about how to *simulate* one, that is, how to use a computer or a table of random numbers to generate a typical "sample

*Unless stated otherwise, when we use the term "Markov chain," we will be restricting our attention to the subclass of *time-homogeneous* Markov chains. We'll do this to avoid monotonous repetition of the phrase "time-homogeneous." I'll point out below the place at which the assumption of time-homogeneity enters.

path." To start, how do I tell you which particular Markov chain I want you to simulate? There are three items involved: to specify a Markov chain, I need to tell you its

- State space $\mathcal{S}$.

  $\mathcal{S}$ is a finite or countable set of *states*, that is, values that the random variables X_i may take on. For definiteness, and without loss of generality, let us label the states as follows: either $\mathcal{S} = \{1, 2, \ldots, N\}$ for some finite N, or $\mathcal{S} = \{1, 2, \ldots\}$, which we may think of as the case "$N = \infty$".

- Initial distribution π_0.

  This is the probability distribution of the Markov chain at time 0. For each state $i \in \mathcal{S}$, we denote by $\pi_0(i)$ the probability $\mathbb{P}\{X_0 = i\}$ that the Markov chain starts out in state i. Formally, π_0 is a function taking $\mathcal{S}$ into the interval [0,1] such that

  $$\pi_0(i) \geq 0 \text{ for all } i \in \mathcal{S}$$

  and

  $$\sum_{i \in \mathcal{S}} \pi_0(i) = 1.$$

  Equivalently, instead of thinking of π_0 as a function from $\mathcal{S}$ to [0,1], we could think of π_0 as the vector whose ith entry is $\pi_0(i) = \mathbb{P}\{X_0 = i\}$.

- Probability transition rule

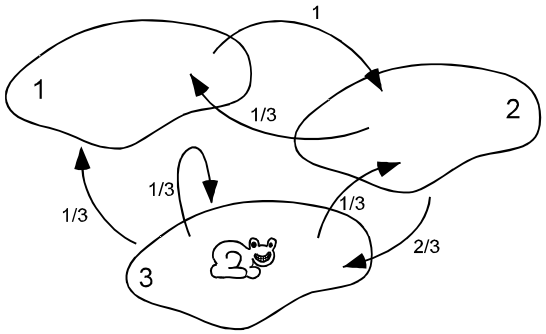
  This is specified by giving a matrix $P = (P_{ij})$. If $\mathcal{S}$ is the finite set $\{1, \ldots, N\}$, say, then P is an $N \times N$ matrix. Otherwise, P will have infinitely many rows and columns; sorry. The interpretation of the number P_{ij} is the conditional probability, given that the chain is in state i at time n, say, that the chain jumps to the state j at time $n+1$. That is,

  $$P_{ij} = \mathbb{P}\{X_{n+1} = j \mid X_n = i\}.$$

  We will also use the notation $P(i, j)$ for the same thing. Note that we have written this probability as a function of just i and j, but of course it could depend on n as well. The ***time homogeneity*** restriction mentioned in the previous footnote is just the assumption that this probability does not depend on the time n, but rather remains constant over time.

  Formally, a ***probability transition matrix*** is an $N \times N$ matrix whose entries are all nonnegative and whose rows sum to 1.

  Finally, you may be wondering why we bother to arrange these conditional probabilities into a matrix. That is a good question, and will be answered soon.



(1.1) FIGURE. *The Markov frog.*

We can now get to the question of how to simulate a Markov chain, now that we know how to specify what Markov chain we wish to simulate. Let's do an example: suppose the state space is $\mathcal{S} = \{1, 2, 3\}$, the initial distribution is $\pi_0 = (1/2, 1/4, 1/4)$, and the probability transition matrix is

$$P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 \\ 1/3 & 0 & 2/3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix} \end{array}. \tag{1.2}$$

Think of a frog hopping among lily pads as in Figure 1.1. How does the Markov frog choose a path? To start, he chooses his initial position X_0 according to the specified initial distribution π_0. He could do this by going to his computer to generate a uniformly distributed random number $U_0 \sim \text{Unif}(0, 1)$, and then taking

$$X_0 = \begin{cases} 1 & \text{if } 0 < U_0 < 1/2 \\ 2 & \text{if } 1/2 < U_0 < 3/4 \\ 3 & \text{if } 3/4 < U_0 < 1 \end{cases}$$

⟦We don't have to be fastidious about specifying what to do if U_0 comes out be exactly $1/2$ or $3/4$, since the probability of this happening is 0.⟧ For example, suppose that U_0 comes out to be 0.8419, so that $X_0 = 3$. Then the frog chooses X_1 according to the probability distribution in row 3 of P, namely, $(1/3, 1/3, 1/3)$; to do this, he paws his computer again to generate $U_1 \sim \text{Unif}(0, 1)$ independently of U_0, and takes

$$X_1 = \begin{cases} 1 & \text{if } 0 < U_0 < 1/3 \\ 2 & \text{if } 1/3 < U_0 < 2/3 \\ 3 & \text{if } 2/3 < U_0 < 1. \end{cases}$$

Suppose he happens to get $U_1 = 0.1234$, so that $X_1 = 1$. Then he chooses X_2 according to row 1 of P, so that $X_2 = 2$; there's no choice this time. Next, he chooses X_3 according to row 2 of P. And so on....

1.2 The Markov property

Clearly, in the previous example, if I told you that we came up with the values $X_0 = 3$, $X_1 = 1$, and $X_2 = 2$, then the conditional probability distribution for X_3 is

$$\mathbb{P}\{X_3 = j \mid X_0 = 3, X_1 = 1, X_2 = 2\} = \begin{cases} 1/3 & \text{for } j = 1 \\ 0 & \text{for } j = 2 \\ 2/3 & \text{for } j = 3, \end{cases}$$

which is also the conditional probability distribution for X_3 given only the information that $X_2 = 2$. In other words, given that $X_0 = 3$, $X_1 = 1$, and $X_2 = 2$, the only information relevant to the distribution to X_3 is the information that $X_2 = 2$; we may ignore the information that $X_0 = 3$ and $X_1 = 1$. This is clear from the description of how to simulate the chain! Thus,

$$\mathbb{P}\{X_3 = j \mid X_2 = 2, X_1 = 1, X_0 = 3\} = \mathbb{P}\{X_3 = j \mid X_2 = 2\} \text{ for all } j.$$

This is an example of the Markov property.

(1.3) DEFINITION. *A process $X_0, X_1, \ldots$ satisfies the* ***Markov property*** *if*

$$\begin{aligned} &\mathbb{P}\{X_{n+1} = i_{n+1} \mid X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\} \\ &\quad = \mathbb{P}\{X_{n+1} = i_{n+1} \mid X_n = i_n\} \end{aligned}$$

for all n and all $i_0, \ldots, i_{n+1} \in \mathcal{S}$.

The issue addressed by the Markov property is the *dependence structure* among random variables. The simplest dependence structure for $X_0, X_1, \ldots$ is no dependence at all, that is, independence. The Markov property could be said to capture the next simplest sort of dependence: in generating the process $X_0, X_1, \ldots$ sequentially, each X_n depends only on the preceding random variable X_{n-1}, and not on the further past values $X_0, \ldots, X_{n-2}$. The Markov property allows much more interesting and general processes to be considered than if we restricted ourselves to independent random variables X_i, without allowing so much generality that a mathematical treatment becomes intractable.

The Markov property implies a simple expression for the probability of our Markov chain taking any specified path, as follows:

$$\begin{aligned} &\mathbb{P}\{X_0 = i_0, X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n\} \\ &= \mathbb{P}\{X_0 = i_0\}\mathbb{P}\{X_1 = i_1 \mid X_0 = i_0\}\mathbb{P}\{X_2 = i_2 \mid X_1 = i_1, X_0 = i_0\} \\ &\qquad \cdots \mathbb{P}\{X_n = i_n \mid X_{n-1} = i_{n-1}, \ldots, X_1 = i_1, X_0 = i_0\} \\ &= \mathbb{P}\{X_0 = i_0\}\mathbb{P}\{X_1 = i_1 \mid X_0 = i_0\}\mathbb{P}\{X_2 = i_2 \mid X_1 = i_1\} \\ &\qquad \cdots \mathbb{P}\{X_n = i_n \mid X_{n-1} = i_{n-1}\} \\ &= \pi_0(i_0)P(i_0, i_1)P(i_1, i_2)\cdots P(i_{n-1}, i_n). \end{aligned}$$

So, to get the probability of a path, we start out with the initial probability of the first state and successively multiply by the matrix elements corresponding to the transitions along the path.

(1.4) EXERCISE. *Let $X_0, X_1, \ldots$ be a Markov chain, and let A and B be subsets of the state space.*

1. *Is it true that $\mathbb{P}\{X_2 \in B \mid X_1 = x_1, X_0 \in A\} = \mathbb{P}\{X_2 \in B \mid X_1 = x_1\}$? Give a proof or counterexample.*
2. *Is it true that $\mathbb{P}\{X_2 \in B \mid X_1 \in A, X_0 = x_0\} = \mathbb{P}\{X_2 \in B \mid X_1 \in A\}$? Give a proof or counterexample.*

[[The moral: be careful about what the Markov property says!]]

(1.5) EXERCISE. *Let $X_0, X_1, \ldots$ be a Markov chain on the state space $\{-1, 0, 1\}$, and suppose that $P(i, j) > 0$ for all i, j. What is a necessary and sufficient condition for the sequence of absolute values $|X_0|, |X_1|, \ldots$ to be a Markov chain?*

(1.6) DEFINITION. *We say that a process $X_0, X_1, \ldots$ is **rth order Markov** if*

$$\begin{aligned}&\mathbb{P}\{X_{n+1} = i_{n+1} \mid X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\}\\ &\quad = \mathbb{P}\{X_{n+1} = i_{n+1} \mid X_n = i_n, \ldots, X_{n-r+1} = i_{n-r+1}\}\end{aligned}$$

for all $n \geq r$ and all $i_0, \ldots, i_{n+1} \in \mathcal{S}$.

(1.7) EXERCISE [A MOVING AVERAGE PROCESS]. *Moving average models are used frequently in time series analysis, economics and engineering. For these models, one assumes that there is an underlying, unobserved process $\ldots, Y_{-1}, Y_0, Y_1, \ldots$ of iid random variables. A **moving average process** takes an average (possibly a weighted average) of these iid random variables in a "sliding window." For example, suppose that at time n we simply take the average of the Y_n and Y_{n-1}, defining $X_n = (1/2)(Y_n + Y_{n-1})$. Our goal is to show that the process $X_0, X_1, \ldots$ defined in this way is not Markov. As a simple example, suppose that the distribution of the iid Y random variables is $\mathbb{P}\{Y_i = 1\} = 1/2 = \mathbb{P}\{Y_i = -1\}$.*

1. *Show that $X_0, X_1, \ldots$ is not a Markov chain.*
2. *Show that $X_0, X_1, \ldots$ is not an rth order Markov chain for any finite r.*

(1.8) NOTATION. *We will use the shorthand "$\mathbb{P}_i$" to indicate a probability taken in a Markov chain started in state i at time 0. That is, "$\mathbb{P}_i(A)$" is shorthand for "$\mathbb{P}\{A \mid X_0 = i\}$." We'll also use the notation "$\mathbb{E}_i$" in an analogous way for expectation.*

(1.9) EXERCISE. *Let $\{X_n\}$ be a finite-state Markov chain and let A be a subset of the state space. Suppose we want to determine the expected time until the chain enters the set A, starting from an arbitrary initial state. That is, letting $\tau_A = \inf\{n \geq 0 : X_n \in A\}$ denote the first time to hit A [[defined to be 0 if $X_0 \in A$]], we want to determine $\mathbb{E}_i(\tau_A)$. Show that*

$$\mathbb{E}_i(\tau_A) = 1 + \sum_k P(i,k)\mathbb{E}_k(\tau_A)$$

for $i \notin A$.

(1.10) EXERCISE. *You are flipping a coin repeatedly. Which pattern would you expect to see faster: HH or HT? For example, if you get the sequence TTHHHTH..., then you see "HH" at the 4th toss and "HT" at the 6th. Letting N_1 and N_2 denote the times required to see "HH" and "HT", respectively, can you guess intuitively whether $\mathbb{E}(N_1)$ is smaller than, the same as, or larger than $\mathbb{E}(N_2)$? Go ahead, make a guess [[and my day]]. Why don't you also simulate some to see how the answer looks; I recommend a computer, but if you like tossing real coins, enjoy yourself by all means. Finally, you can use the reasoning of the Exercise (1.9) to solve the problem and evaluate $\mathbb{E}(N_i)$. A hint is to set up a Markov chain having the 4 states HH, HT, TH, and TT.*

(1.11) EXERCISE. *Here is a chance to practice formalizing some typical "intuitively obvious" statements. Let $X_0, X_1, \ldots$ be a finite-state Markov chain.*

a. *We start with an observation about conditional probabilities that will be a useful tool throughout the rest of this problem. Let $F_1, \ldots, F_m$ be disjoint events. Show that if $\mathbb{P}(E|F_i) = p$ for all $i = 1, \ldots, m$ then $\mathbb{P}(E \mid \bigcup_{i=1}^m F_i) = p$.*

b. *Show that*

$$\begin{aligned}&\mathbb{P}\{X_{n+1} \in A_1, \ldots, X_{n+r} \in A_r \mid X_n = j, X_{n-1} \in B_{n-1}, \ldots, X_0 \in B_0\}\\ &\quad = \mathbb{P}_j\{X_{n+1} \in A_1, \ldots, X_{n+r} \in A_r\}.\end{aligned}$$

c. *Recall the definition of hitting times: $T_i = \inf\{n > 0 : X_n = i\}$. Show that $\mathbb{P}_i\{T_i = n + m \mid T_j = n, T_i > n\} = \mathbb{P}_j\{T_i = m\}$, and conclude that $\mathbb{P}_i\{T_i = T_j + m \mid T_j < \infty, T_i > T_j\} = \mathbb{P}_j\{T_i = m\}$. This is one manifestation of the statement that the Markov chain "probabilistically restarts" after it hits j.*

d. *Show that $\mathbb{P}_i\{T_i < \infty \mid T_j < \infty, T_i > T_j\} = \mathbb{P}_j\{T_i < \infty\}$. Use this to show that if $\mathbb{P}_i\{T_j < \infty\} = 1$ and $\mathbb{P}_j\{T_i < \infty\} = 1$, then $\mathbb{P}_i\{T_i < \infty\} = 1$.*

e. *Let i be a recurrent state and let $j \neq i$. Recall the idea of "cycles," the segments of the path between successive visits to i. For simplicity let's just look at the first two cycles. Formulate and prove an assertion to the effect that whether or not the chain visits state j during the first and second cycles can be described by iid Bernoulli random variables.*

1.3 "It's all just matrix theory"

Recall that the vector π_0 having components $\pi_0(i) = \mathbb{P}\{X_0 = i\}$ is the initial distribution of the chain. Let π_n denote the distribution of the chain at time n, that is, $\pi_n(i) = \mathbb{P}\{X_n = i\}$. Suppose for simplicity that the state space is finite: $\mathcal{S} = \{1, \ldots, N\}$, say. Then the Markov chain has an $N \times N$ probability transition matrix

$$P = (P_{ij}) = (P(i,j)),$$

where $P(i,j) = \mathbb{P}\{X_{n+1} = j \mid X_n = i\} = \mathbb{P}\{X_1 = j \mid X_0 = i\}$. The law of total probability gives

$$\begin{aligned}\pi_{n+1}(j) &= \mathbb{P}\{X_{n+1} = j\} \\ &= \sum_{i=1}^{N} \mathbb{P}\{X_n = i\}\mathbb{P}\{X_{n+1} = j \mid X_n = i\} \\ &= \sum_{i=1}^{N} \pi_n(i) P(i,j),\end{aligned}$$

which, in matrix notation, is just the equation

$$\pi_{n+1} = \pi_n P.$$

Note that here we are thinking of π_n and π_{n+1} as *row vectors*, so that, for example,

$$\pi_n = (\pi_n(1), \ldots, \pi_n(N)).$$

Thus, we have

$$\begin{aligned}\pi_1 &= \pi_0 P \\ \pi_2 &= \pi_1 P = \pi_0 P^2 \\ \pi_3 &= \pi_2 P = \pi_0 P^3,\end{aligned} \tag{1.12}$$

and so on, so that by induction

$$\pi_n = \pi_0 P^n. \tag{1.13}$$

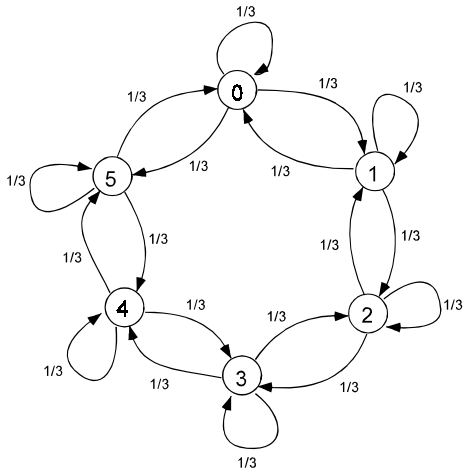
(1.14) EXERCISE. *Let $P^n(i,j)$ denote the (i,j) element in the matrix P^n, the nth power of P. Show that $P^n(i,j) = \mathbb{P}\{X_n = j \mid X_0 = i\}$. Ideally, you should get quite confused about what is being asked, and then straighten it all out.*

So, in principle, we can find the answer to any question about the probabilistic behavior of a Markov chain by doing matrix algebra, finding powers of matrices, etc. However, what is viable in practice may be another story. For example, the state space for a Markov chain that describes repeated shuffling of a deck of cards contains 52! elements—the permutations of the 52 cards of the deck. This number 52! is large: about 80 million million million million

million million million million million million million. The probability transition matrix that describes the effect of a single shuffle is a 52! by 52! matrix. So, "all we have to do" to answer questions about shuffling is to take powers of such a matrix, find its eigenvalues, and so on! In a practical sense, simply reformulating probability questions as matrix calculations often provides only minimal illumination in concrete questions like "how many shuffles are required in order to mix the deck well?" Probabilistic reasoning can lead to insights and results that would be hard to come by from thinking of these problems as "just" matrix theory problems.

1.4 The basic limit theorem of Markov chains

As indicated by its name, the theorem we will discuss in this section occupies a fundamental and important role in Markov chain theory. What is it all about? Let's start with an example in which we can all see intuitively what is going on.



(1.15) FIGURE. *A random walk on a clock.*

(1.16) EXAMPLE [RANDOM WALK ON A CLOCK]. For ease of writing and drawing, consider a clock with 6 numbers on it: 0,1,2,3,4,5. Suppose we perform a random walk by moving clockwise, moving counterclockwise, and staying in place with probabilities 1/3 each at every time n. That is,

$$P(i,j) = \begin{cases} 1/3 & \text{if } j = i-1 \bmod 6 \\ 1/3 & \text{if } j = i \\ 1/3 & \text{if } j = i+1 \bmod 6. \end{cases}$$

Suppose we start out at $X_0 = 2$, say. That is,

$$\pi_0 = (\pi_0(0), \pi_0(1), \ldots, \pi_0(5)) = (0,0,1,0,0,0).$$

Then of course

$$\pi_1 = (0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0),$$

and it is easy to calculate

$$\pi_2 = (\frac{1}{9}, \frac{2}{9}, \frac{1}{3}, \frac{2}{9}, \frac{1}{9}, 0)$$

and

$$\pi_3 = (\frac{3}{27}, \frac{6}{27}, \frac{7}{27}, \frac{6}{27}, \frac{3}{27}, \frac{2}{27}).$$

Notice how the probability is spreading out away from its initial concentration on the state 2. We could keep calculating π_n for more values of n, but it is intuitively clear what will happen: the probability will continue to spread out, and π_n will approach the uniform distribution:

$$\pi_n \to (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$$

as $n \to \infty$. Just imagine: if the chain starts out in state 2 at time 0, then we close our eyes while the random walk takes 10,000 steps, and then we are asked to guess what state the random walk is in at time 10,000, what would we think the probabilities of the various states are? I would say: "$X_{10,000}$ is for all practical purposes uniformly distributed over the 6 states." By time 10,000, the random walk has essentially "forgotten" that it started out in state 2 at time 0, and it is nearly equally likely to be anywhere.

Now observe that the starting state 2 was not special; we could have started from anywhere, and over time the probabilities would spread out away from the initial point, and approach the same limiting distribution. Thus, π_n approaches a limit that does not depend upon the initial distribution π_0. □

The following "Basic Limit Theorem" says that the phenomenon discussed in the previous example happens quite generally. We will start with a statement and discussion of the theorem, and then prove the theorem later. We'll use the notation "$\mathbb{P}_{\pi_0}$" for probabilities when the initial distribution is π_0.

(1.17) THEOREM [BASIC LIMIT THEOREM]. *Let $X_0, X_1, \ldots$ be an irreducible, aperiodic Markov chain having a stationary distribution $\pi(\cdot)$. Then for all initial distributions π_0,*

$$\lim_{n\to\infty} \mathbb{P}_{\pi_0}\{X_n = i\} = \pi(i) \text{ for all } i \in \mathcal{S}.$$

We need to define the words "irreducible," "aperiodic," and "stationary distribution." Let's start with "stationary distribution."

1.5 Stationary distributions

Suppose a distribution π on $\mathcal{S}$ is such that, if our Markov chain starts out with initial distribution $\pi_0 = \pi$, then we also have $\pi_1 = \pi$. That is, if the distribution at time 0 is π, then the distribution at time 1 is still π. Then π is called a ***stationary distribution*** for

the Markov chain. From (1.12) we see that the definition of stationary distribution amounts to saying that π satisfies the equation

$$\pi = \pi P, \tag{1.18}$$

that is,

$$\pi(j) = \sum_{i \in \mathcal{S}} \pi(i) P(i,j) \quad \text{for all } j \in \mathcal{S}.$$

[[In the case of an infinite state space, (1.18) is an infinite system of equations.]] Also from equations (1.12) we can see that if the Markov chain has initial distribution $\pi_0 = \pi$, then we have not only $\pi_1 = \pi$, but also $\pi_n = \pi$ for all n. That is, a Markov chain started out in a stationary distribution π stays in the distribution π forever; that's why the distribution π is called "stationary."

(1.19) EXAMPLE. If the $N \times N$ probability transition matrix P is symmetric, then the uniform distribution [[$\pi(i) = 1/N$ for all i]] is stationary. More generally, the uniform distribution is stationary if the matrix P is *doubly stochastic*, that is, the column-sums of P are 1 (we already know the row-sums of P are all 1). □

It should not be surprising that π appears as the limit in Theorem (1.17). It is easy to see that if π_n approaches a limiting distribution as $n \to \infty$, then that limiting distribution must be stationary. To see this, suppose that $\lim_{n\to\infty} \pi_n = \tilde{\pi}$, and let $n \to \infty$ in the equation $\pi_{n+1} = \pi_n P$ to obtain $\tilde{\pi} = \tilde{\pi} P$, which says that $\tilde{\pi}$ is stationary.

(1.20) EXERCISE [FOR THE MATHEMATICALLY INCLINED]. *What happens in the case of a countably infinite state space? Does the sort of argument in the previous paragraph still work?*

Computing stationary distributions is an algebra problem. Since most people are accustomed to solving linear systems of the form $Ax = b$, let us take the transpose of the equation $\pi(P - I) = 0$, getting the equation $(P^T - I)\pi^T = 0$. For example, for the matrix P from (1.2), we get the equation

$$\begin{pmatrix} -1 & 1/3 & 1/3 \\ 1 & -1 & 1/3 \\ 0 & 2/3 & -2/3 \end{pmatrix} \begin{pmatrix} \pi(1) \\ \pi(2) \\ \pi(3) \end{pmatrix} = 0,$$

or

$$\begin{pmatrix} -1 & 1/3 & 1/3 \\ 0 & -2/3 & 2/3 \\ 0 & 2/3 & -2/3 \end{pmatrix} \begin{pmatrix} \pi(1) \\ \pi(2) \\ \pi(3) \end{pmatrix} = 0,$$

which has solutions of the form $\pi = \text{const}(2/3, 1, 1)$. For the unique solution that satisfies the constraint $\sum \pi(i) = 1$, take the constant to be $3/8$, so that $\pi = (1/4, 3/8, 3/8)$.

Here is another way, aside from solving the linear equations, to approach the problem of finding a stationary distribution; this idea can work particularly well with computers. If

we believe the Basic Limit Theorem, we should see the stationary distribution in the limit as we run the chain for a long time. Let's try it: Here are some calculations of powers of the transition matrix P from (1.2):

$$P^5 = \begin{pmatrix} 0.246914 & 0.407407 & 0.345679 \\ 0.251029 & 0.36214 & 0.386831 \\ 0.251029 & 0.366255 & 0.382716 \end{pmatrix},$$

$$P^{10} = \begin{pmatrix} 0.250013 & 0.37474 & 0.375248 \\ 0.249996 & 0.375095 & 0.374909 \\ 0.249996 & 0.375078 & 0.374926 \end{pmatrix},$$

$$P^{20} = \begin{pmatrix} 0.2500000002 & 0.3749999913 & 0.3750000085 \\ 0.2499999999 & 0.375000003 & 0.374999997 \\ 0.2499999999 & 0.3750000028 & 0.3749999973 \end{pmatrix}.$$

So we don't really have to solve equations; in this example, any of the rows of the matrix P^{20} provides a very accurate approximation for π. No matter what state we start from, the distribution after 20 steps of the chain is very close to $(.25, .375, .375)$. This is the Basic Limit Theorem in action.

(1.21) EXERCISE [STATIONARY DISTRIBUTION OF EHRENFEST CHAIN]. *The Ehrenfest chain is a simple model of "mixing" processes. This chain can shed light on perplexing questions like "Why aren't people dying all the time due to the air molecules bunching up in some odd corner of their bedrooms while they sleep?" The model considers d balls distributed among two urns, and results in a Markov chain $\{X_0, X_1, \ldots\}$ having state space $\{0, 1, \ldots, d\}$, with the state X_n of the chain at time n being the number of balls in urn #1 at time n. At each time, we choose a ball at random uniformly from the d possibilities, take that ball out of its current urn, and drop it into the other urn. Thus, $P(i, i-1) = i/d$ and $P(i, i+1) = (d-i)/d$ for all i.*

What is the stationary distribution of the Ehrenfest chain? You might want to solve the problem for a few small values of d. You should notice a pattern, and come up with a familiar answer. Can you explain without calculation why this distribution is stationary?

A Markov chain might have no stationary distribution, one stationary distribution, or infinitely many stationary distributions. We just saw an example with one. A trivial example with infinitely many is when P is the identity matrix, in which case all distributions are stationary. To find an example without any stationary distribution, we need to consider an infinite state space. ⟦We will see later that any finite-state Markov chain has at least one stationary distribution.⟧ An easy example of this has $\mathcal{S} = \{1, 2, \ldots\}$ and $P(i, i+1) = 1$ for all i, which corresponds to a Markov chain that moves deterministically "to the right." In this case, the equation $\pi(j) = \sum_{i \in \mathcal{S}} \pi(i) P(i, j)$ reduces to $\pi(j) = \pi(j-1)$, which clearly has no solution satisfying $\sum \pi(j) = 1$. Another interesting example is the *simple, symmetric random walk on the integers*: $P(i, i-1) = 1/2 = P(i, i+1)$. Here the equations for stationarity become

$$\pi(j) = \frac{1}{2}\pi(j-1) + \frac{1}{2}\pi(j+1).$$

Again it is easy to see [[how?]] that these equations have no solution π that is a probability mass function.

Intuitively, notice the qualitative difference: in the examples without a stationary distribution, the probability doesn't settle down to a limit probability distribution—in the first example the probability moves off to infinity, and in the second example it spreads out in both directions. In both cases, the probability on any fixed state converges to 0; one might say the probability escapes off to infinity (or $-\infty$). How can we keep the probability from escaping? Here is an example.

(1.22) EXERCISE. *Consider a Markov chain on the integers with*

$$P(i,i+1) = .4 \text{ and } P(i,i-1) = .6 \text{ for } i > 0,$$
$$P(i,i+1) = .6 \text{ and } P(i,i-1) = .4 \text{ for } i < 0,$$
$$P(0,1) = P(0,-1) = 1/2.$$

This is a chain with infinitely many states, but it has a sort of probabilistic "restoring force" that always pushes back toward 0. Find the stationary distribution.

The next exercise may look a bit inscrutable at first, but it is well worth doing and it introduces an important idea.

(1.23) EXERCISE [PROBABILITY FLUX]. *Consider a partition of the state space $\mathcal{S}$ of a Markov chain into two subsets A and A^c. Suppose the Markov chain has stationary distribution π. Show that*

(1.24)
$$\sum_{i\in A}\sum_{j\in A^c} \pi(i)P(i,j) = \sum_{i\in A^c}\sum_{j\in A} \pi(i)P(i,j).$$

(1.25) EXERCISE. *Use exercise (1.23) to re-do Exercise (1.21), by writing the equations produced by (1.24) with the choice $A = \{0,1,\ldots,i\}$ for various i. The calculation should be easier.*

The left side of (1.24) may be thought of the "probability flux flowing out of A into A^c." The equality says that this must be the same as the flux from A^c back into A. This has the suggestive interpretation that the stationary probabilities describe a stable system in which all the probability is happy where it is, and does not want to flow to anywhere else, so that the net flow from A to A^c must be zero. We can say this in a much less mysterious way as follows. Think of $\pi(i)$ as the long run fraction of time that the chain is in state i. [[We will soon see a theorem ("a strong law of large numbers for Markov chains") that supports this interpretation.]] Then $\pi(i)P(i,j)$ is the long run fraction of times that a transition from i to j takes place. But clearly the long run fraction of times occupied by transitions going from a state in A to a state in A^c must equal the long run fraction of times occupied by transitions going the opposite way. [[In fact, along any sample path, the numbers of

transitions that have occurred in the two directions up to any time n may differ by at most 1!⟧

(1.26) EXERCISE [RENEWAL THEORY, THE RESIDUAL, AND LENGTH-BIASED SAMPLING]. *Let $X_1, X_2, \ldots$ be iid taking values in $\{1, \ldots, d\}$. ⟦These are typically thought of as lifetimes of lightbulbs...⟧ Define $S_k = X_1 + \cdots + X_k$, $\tau(n) = \inf\{k : S_k \geq n\}$, and $R_n = S_{\tau(n)} - n$. Then R_n is called the* residual lifetime *at time n. ⟦This is the amount of lifetime remaining in the bulb that is in operation at time n.⟧*

1. *The sequence $R_0, R_1, \ldots$ is a Markov chain. What is its transition matrix? What is the stationary distribution?*

2. *Define the* total lifetime *L_n at time n by $L_n = X_{\tau(n)}$. This has an obvious interpretation as the total lifetime of the lightbulb in operation at time n. Show that $L_0, L_1, \ldots$ is not a Markov chain. But L_n still has a limiting distribution, and we'd like to find it. We'll do this by constructing a Markov chain by enlarging the state space and considering the sequence of random vectors $(R_0, L_0), (R_1, L_1), \ldots$. This sequence does form a Markov chain. What is its probability transition function and stationary distribution? Now, assuming the Basic Limit Theorem applies here, what is the limiting distribution of L_n as $n \to \infty$? This is the famous "length-biased sampling" distribution.*

1.6 Irreducibility, periodicity, and recurrence

We now turn to the definition of irreducibility. Let i and j be two states. We say that j ***is accessible from*** i if it is possible ⟦with positive probability⟧ for the chain ever to visit state j if the chain starts in state i, or, in other words,

$$\mathbb{P}\{\bigcup_{n=0}^{\infty}\{X_n = j\} \mid X_0 = i\} > 0.$$

Clearly an equivalent condition is

$$\sum_{n=0}^{\infty} P^n(i,j) \stackrel{\triangle}{=} \sum_{n=0}^{\infty} \mathbb{P}\{X_n = j \mid X_0 = i\} > 0. \tag{1.27}$$

(1.28) EXERCISE. *Prove the last assertion.*

We say i ***communicates with*** j if j is accessible from i and i is accessible from j.

(1.29) EXERCISE. *Show that the relation "communicates with" is an equivalence relation. That is, show that the "communicates with" relation is reflexive, symmetric, and transitive.*

We say that the Markov chain is ***irreducible*** if all pairs of states communicate.

Recall that an equivalence relation on a set induces a partition of that set into equivalence classes. Thus, by Exercise (1.29), the state space $\mathcal{S}$ may be partitioned into what we will call "communicating classes," or simply "classes." The chain is irreducible if there is just one communicating class, that is, the whole state space $\mathcal{S}$. Note that whether or not a Markov chain is irreducible is determined by the state space $\mathcal{S}$ and the transition matrix $(P(i,j))$; the initial distribution π_0 is irrelevant. In fact, all that matters is the pattern of zeroes in the transition matrix.

Why do we require irreducibility in the "Basic Limit Theorem" (1.17)? Here is a trivial example of how the conclusion can fail if we do not assume irreducibility. Let $\mathcal{S} = \{0,1\}$ and let $P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Clearly the resulting Markov chain is not irreducible. Also, clearly the conclusion of the Basic Limit Theorem does not hold; that is, π_n does not approach any limit that is independent of π_0. In fact, $\pi_n = \pi_0$ for all n.

Next, to discuss periodicity, let's begin with another trivial example: take $\mathcal{S} = \{0,1\}$ again, and let $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. The conclusion of the Basic Limit Theorem does not hold here: for example, if $\pi_0 = (1,0)$, then $\pi_n = (1,0)$ if n is even and $\pi_n = (0,1)$ if n is odd. So in this case $\pi_n(1)$ alternates between the two values 0 and 1 as n increases, and hence does not converge to anything. The problem in this example is not lack of irreducibility; clearly this chain is irreducible. So, assuming the Basic Limit Theorem is true, the chain must not be aperiodic! That is, the chain is ***periodic***. The trouble stems from the fact that, starting from state 1 at time 0, the chain can visit state 1 only at even times. The same holds for state 2.

Given a Markov chain $\{X_0, X_1, \ldots\}$, define the ***period*** of a state i to be

$$d_i = \gcd\{n : P^n(i,i) > 0\}.$$

Note that both states 1 and 2 in the example $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ have period 2. In fact, the next result shows that if two states i and j communicate, then they must have the same period.

(1.30) THEOREM. *If the states i and j communicate, then $d_i = d_j$.*

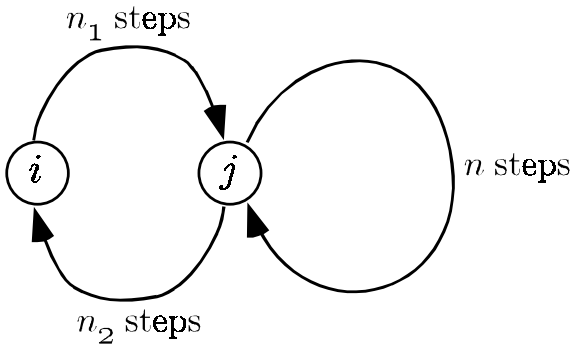
PROOF: Since j is accessible from i, by (1.27) there exists an n_1 such that $P^{n_1}(i,j) > 0$. Similarly, since i is accessible from j, there is an n_2 such that $P^{n_2}(j,i) > 0$. Noting that $P_{n_1+n_2}(i,i) > 0$, it follows that

$$d_i \mid n_1 + n_2,$$

that is, d_i divides $n_1 + n_2$, which means that $n_1 + n_2$ is an integer multiple of d_i. Now suppose that $P^n(j,j) > 0$. Then $P^{n_1+n+n_2}(i,i) > 0$, so that

$$d_i \mid n_1 + n + n_2.$$

Subtracting the last two displays gives $d_i \mid n$. Since n was an arbitrary integer satisfying $P^n(j,j) > 0$, we have found that d_i is a common divisor of the set $\{n : P^n(j,j) > 0\}$. Since d_j is defined to be the *greatest* common divisor of this set, we have shown that $d_j \geq d_i$. Interchanging the roles of i and j in the previous argument gives the opposite inequality $d_i \geq d_j$. This completes the proof. □



It follows from Theorem (1.30) that all states in a communicating class have the same period. We say that the period of a state is a "class property." In particular, all states in an irreducible Markov chain have the same period. Thus, we can speak of *the period of a Markov chain* if that Markov chain is irreducible: the period of an irreducible Markov chain is the period of any of its states.

(1.31) DEFINITION. *An irreducible Markov chain is said to be* ***aperiodic*** *if its period is 1, and* ***periodic*** *otherwise.*

We have now discussed all of the words we need in order to understand the statement of the Basic Limit Theorem (1.17). We will need another concept or two before we can get to the proof, and the proof will then take some time beyond that. So I propose that we pause to discuss an interesting example of an application of the Basic Limit Theorem; this will help us build up some motivation to help carry us through the proof, and will also give some practice that should help be helpful in assimilating the concepts of irreducibility and aperiodicity.

(1.32) EXAMPLE [GENERATING A RANDOM TABLE WITH FIXED ROW AND COLUMN SUMS]. Consider the 4×4 table of numbers that is enclosed within the rectangle below. The four numbers along the bottom of the table are the column sums, and those along the right edge of the table are the row sums.

68	119	26	7	220
20	84	17	94	215
15	54	14	10	93
5	29	14	16	64
108	286	71	127	

Suppose we want to generate a random 4×4 table that has the same row and column sums as the table above. That is, suppose that we want to generate a random table of nonnegative integers whose probability distribution is uniform on the set $\mathcal{S}$ of all such 4×4 tables that have the given row and column sums. Here is a proposed algorithm. Start with any table having the correct row and column sums; so of course the 4×4 table given above will do. Denote the entries in that table by a_{ij}. Choose a pair $\{i_1, i_2\}$ of rows at random, that is, uniformly over the $\binom{4}{2} = 6$ possible pairs. Similarly, choose a random pair of columns $\{j_1, j_2\}$. Then flip a coin. If you get heads: add 1 to $a_{i_1 j_1}$ and $a_{i_2 j_2}$, and subtract 1 from $a_{i_1 j_2}$ and $a_{i_2 j_1}$ *if you can do so without producing any negative entries*—if you cannot do so, then do nothing. Similarly, if the coin flip comes up tails, then subtract 1 from $a_{i_1 j_1}$ and $a_{i_2 j_2}$, and add 1 to $a_{i_1 j_2}$ and $a_{i_2 j_1}$, with the same nonnegativity proviso, and otherwise do nothing. This describes a random transformation of the original table that results in a new table in the desired set of tables $\mathcal{S}$. Now repeat the same random transformation on the new table, and so on. □

(1.33) EXERCISE. *Assuming the validity of the Basic Limit Theorem, show that if we run the "algorithm" in Example (1.32) for "a long time," then we will end up with a random table having probability distribution very close to the desired distribution. In order to do this, show that*

1. *The procedure generates a Markov chain whose state space is* $\mathcal{S}$,
2. *that Markov chain is irreducible,*
3. *that Markov chain is aperiodic, and*
4. *that Markov chain has the desired distribution (that is, uniform on* $\mathcal{S}$*) as its stationary distribution.*

I consider Exercise (1.33) to be an interesting application of the Basic Limit Theorem. I hope it helps whet your appetite for digesting the proof of that theorem!

For the proof of the Basic Limit Theorem, we will need one more concept: *recurrence*. Analogously to what we did with the notion of periodicity, we will begin by saying what a recurrent state is, and then show ⟦in Theorem (1.35) below⟧ that recurrence is actually a class property. In particular, in an irreducible Markov chain, either all states are recurrent or all states are *transient*, which means "not recurrent." Thus, if a chain is irreducible, we can speak of the chain being either recurrent or transient.

The idea of recurrence is this: a state i is recurrent if, starting from the state i at time 0, the chain is sure to return to i eventually. More precisely, define the *first hitting time* T_i of the state i by

$$T_i = \inf\{n > 0 : X_n = i\},$$

and make the following definition.

(1.34) DEFINITION. *The state i is* ***recurrent*** *if $\mathbb{P}_i\{T_i < \infty\} = 1$. If i is not recurrent, it is called* ***transient****.*

The meaning of recurrence is this: state i is recurrent if, when the Markov chain is started out in state i, the chain is *certain* to return to i at some finite future time. Observe the difference in spirit between this and the definition of "accessible from" ⟦see the paragraph containing (1.27)⟧, which requires only that it be *possible* for the chain to hit a state j. In terms of the first hitting time notation, the definition of "accessible from" may be restated as follows: for distinct states $i \neq j$, we say that j is accessible from i if and only if $\mathbb{P}_i\{T_j < \infty\} > 0$. ⟦Why did I bother to say "for distinct states $i \neq j$"?⟧

Here is the promised result that implies that recurrence is a class property.

(1.35) THEOREM. *Let i be a recurrent state, and suppose that j is accessible from i. Then in fact all of the following hold:*

(i) $\mathbb{P}_i\{T_j < \infty\} = 1$;

(ii) $\mathbb{P}_j\{T_i < \infty\} = 1$;

(iii) *The state j is recurrent.*

PROOF: The proof will be given somewhat informally; it can be rigorized. Suppose $i \neq j$, since the result is trivial otherwise.

Firstly, let us observe that (iii) follows from (i) and (ii): clearly if (ii) holds ⟦that is, starting from j the chain is certain to visit i eventually⟧ and (i) holds ⟦so that starting from i the chain is certain to visit j eventually⟧, then (iii) must also hold ⟦since starting from j the chain is certain to visit i, after which it will definitely get back to j⟧.

To prove (i), let us imagine starting the chain in state i, so that $X_0 = i$. With probability one, the chain returns at some time $T_i < \infty$ to i. For the same reason, continuing the chain after time T_i, the chain is sure to return to i for a second time. In fact, by continuing this argument we see that, with probability one, the chain returns to i infinitely many times. Thus, we may visualize the path followed by the Markov chain as a succession of infinitely many "cycles," where a cycle is a portion of the path between two successive visits to i. That is, we'll say that the first cycle is the segment $X_1, \ldots, X_{T_i}$ of the path, the second cycle starts with X_{T_i+1} and continues up to and including the second return to i, and so on. The behaviors of the chain in successive cycles are independent and have identical probabilistic characteristics. In particular, letting $I_n = 1$ if the chain visits j sometime during the nth cycle and $I_n = 0$ otherwise, we see that $I_1, I_2, \ldots$ is an *iid* sequence of Bernoulli trials. Let p denote the common "success probability"

$$p = \mathbb{P}\{\text{visit } j \text{ in a cycle}\} = \mathbb{P}_i\left[\bigcup_{k=1}^{T_i}\{X_k = j\}\right]$$

for these trials. Clearly if p were 0, then with probability one the chain would not visit j in any cycle, which would contradict the assumption that j is accessible from i. Therefore,

$p > 0$. Now observe that in such a sequence of *iid* Bernoulli trials with a positive success probability, with probability one we will eventually observe a success. In fact,

$$\mathbb{P}_i\{\text{chain does not visit } j \text{ in the first } n \text{ cycles}\} = (1-p)^n \to 0$$

as $n \to \infty$. That is, with probability one, eventually there will be a cycle in which the chain does visit j, so that (i) holds.

It is also easy to see that (ii) must hold. In fact, suppose to the contrary that $\mathbb{P}_j\{T_i = \infty\} > 0$. Combining this with the hypothesis that j is accessible from i, we see that it is possible with positive probability for the chain to go from i to j in some finite amount of time, and then, continuing from state j, never to return to i. But this contradicts the fact that starting from i the chain must return to i infinitely many times with probability one. Thus, (ii) holds, and we are done. □

The "cycle" idea used in the previous proof is powerful and important; we will be using it again.

The next theorem gives a useful equivalent condition for recurrence. The statement uses the notation N_i for the total number of visits of the Markov chain to the state i, that is,

$$N_i = \sum_{n=0}^{\infty} I\{X_n = i\}.$$

(1.36) THEOREM. *The state i is recurrent if and only if $\mathbb{E}_i(N_i) = \infty$.*

PROOF: We already know that if i is recurrent, then

$$\mathbb{P}_i\{N_i = \infty\} = 1,$$

that is, starting from i, the chain visits i infinitely many times with probability one. But of course the last display implies that $\mathbb{E}_i(N_i) = \infty$. To prove the converse, suppose that i is transient, so that $q := \mathbb{P}_i\{T_i = \infty\} > 0$. Considering the sample path of the Markov chain as a succession of "cycles" as in the proof of Theorem (1.35), we see that each cycle has probability q of never ending, so that there are no more cycles, and no more visits to i. In fact, a bit of thought shows that N_i, the total number of visits to i [[including the visit at time 0]], has a geometric distribution with "success probability" q, and hence expected value $1/q$, which is finite, since $q > 0$. □

(1.37) COROLLARY. *If j is transient, then $\lim_{n\to\infty} P^n(i,j) = 0$ for all states i.*

PROOF: Supposing j is transient, we know that $\mathbb{E}_j(N_j) < \infty$. Starting from an arbitrary state $i \neq j$, we have

$$\mathbb{E}_i(N_j) = \mathbb{P}_i\{T_j < \infty\}\mathbb{E}_i(N_j \mid T_j < \infty).$$

However, $\mathbb{E}_i(N_j \mid T_j < \infty) = \mathbb{E}_j(N_j)$; this is clear intuitively since, starting from i, if the Markov chain hits j at the finite time T_j, then it "probabilistically restarts" at time T_j. [[Exercise: give a formal argument.]] Thus, $\mathbb{E}_i(N_j) \leq \mathbb{E}_j(N_j) < \infty$, so that in fact we have $\mathbb{E}_i(N_j) = \sum_{n=1}^{\infty} P^n(i,j) < \infty$, which implies the conclusion of the Corollary. □

(1.38) EXAMPLE ["A DRUNK MAN WILL FIND HIS WAY HOME, BUT A DRUNK BIRD MAY GET LOST FOREVER," OR, RECURRENCE AND TRANSIENCE OF RANDOM WALKS]. The quotation is from Yale's own professor Kakutani, as told by R. Durrett in his probability book. We'll consider a certain model of a random walk in d dimensions, and show that the walk is recurrent if $d = 1$ or $d = 2$, and the walk is transient if $d \geq 3$.

In one dimension, our random walk is the "simple, symmetric" random walk on the integers, which takes steps of $+1$ and -1 with probability $1/2$ each. That is, letting $X_1, X_2, \ldots$ be *iid* taking the values ± 1 with probability $1/2$, we define the position of the random walk at time n to be $S_n = X_1 + \cdots + X_n$. What is a random walk in d dimensions? Here is what we will take it to be: the position of such a random walk at time n is

$$S_n = (S_n(1), \ldots, S_n(d)) \in \mathbb{Z}^d,$$

where the coordinates $S_n(1), \ldots, S_n(d)$ are independent simple, symmetric random walks in $\mathbb{Z}$. That is, to form a random walk in $\mathbb{Z}^d$, simply concatenate d independent one-dimensional random walks into a d-dimensional vector process.

Thus, our random walk S_n may be written as $S_n = X_1 + \cdots + X_n$, where $X_1, X_2, \ldots$ are *iid* taking on the 2^d values $(\pm 1, \ldots, \pm 1)$ with probability 2^{-d} each. This might not be the first model that would come to your mind. Another natural model would be to have the random walk take a step by choosing one of the d coordinate directions at random (probability $1/d$ each) and then taking a step of $+1$ or -1 with probability $1/2$. That is, the increments $X_1, X_2, \ldots$ would be *iid* taking the $2d$ values

$$(\pm 1, 0, \ldots, 0), (0, \pm 1, \ldots, 0), \ldots, (0, 0, \ldots, \pm 1)$$

with probability $1/2d$ each. This is indeed a popular model, and can be analyzed to reach the conclusion "recurrent in $d \leq 2$ and transient in $d \geq 3$" as well. But the "concatenation of d independent random walks" model we will consider is a bit simpler to analyze. Also, for all you Brownian motion fans out there, our model is the random walk analog of d-dimensional Brownian motion, which is a concatenation of d independent one-dimensional Brownian motions.

We'll start with $d = 1$. It is obvious that $S_0, S_1, \ldots$ is an irreducible Markov chain. Since recurrence is a class property, to show that every state is recurrent it suffices to show that the state 0 is recurrent. Thus, by Theorem (1.36) we want to show that

$$\mathbb{E}_0(N_0) = \sum_n P^n(0,0) = \infty. \tag{1.39}$$

But $P^n(0,0) = 0$ if n is odd, and for even $n = 2m$, say, $P^{2m}(0,0)$ is the probability that a Binomial$(2m, 1/2)$ takes the value m, or

$$P^{2m}(0,0) = \binom{2m}{m} 2^{-2m}.$$

This can be closely approximated in a convenient form by using Stirling's formula, which says that

$$k! \sim \sqrt{2\pi k}\,(k/e)^k,$$

where the notation "$a_k \sim b_k$" means that $a_k/b_k \to 1$ as $k \to \infty$. Applying Stirling's formula gives

$$P^{2m}(0,0) = \frac{(2m)!}{(m!)^2 2^{2m}} \sim \frac{\sqrt{2\pi(2m)}\,(2m/e)^{2m}}{2\pi m(m/e)^{2m}2^{2m}} = \frac{1}{\sqrt{\pi m}}.$$

Thus, from the fact that $\sum(1/\sqrt{m}) = \infty$ it follows that (1.39) holds, so that the random walk is recurrent.

Now it's easy to see what happens in higher dimensions. In $d = 2$ dimensions, for example, again we have an irreducible Markov chain, so we may determine the recurrence or transience of chain by determining whether the sum

$$\sum_{n=0}^{\infty} \mathbb{P}_{(0,0)}\{S_{2n} = (0,0)\} \tag{1.40}$$

is infinite or finite, where S_{2n} is the vector (S^1_{2n}, S^2_{2n}), say. By the assumed independence of the two components of the random walk, we have

$$\mathbb{P}_{(0,0)}\{S_{2m} = (0,0)\} = \mathbb{P}_0\{S^1_{2m} = 0\}\mathbb{P}_0\{S^2_{2m} = 0\} \sim \left(\frac{1}{\sqrt{\pi m}}\right)\left(\frac{1}{\sqrt{\pi m}}\right) = \frac{1}{\pi m},$$

so that (1.40) is infinite, and the random walk is again recurrent. However, in $d = 3$ dimensions, the analogous sum

$$\sum_{n=0}^{\infty} \mathbb{P}_{(0,0,0)}\{S_{2n} = (0,0,0)\}$$

is finite, since

$$\mathbb{P}_{(0,0,0)}\{S_{2m} = (0,0,0)\} = \mathbb{P}_0\{S^1_{2m} = 0\}\mathbb{P}_0\{S^2_{2m} = 0\}\mathbb{P}_0\{S^3_{2m} = 0\} \sim \left(\frac{1}{\sqrt{\pi m}}\right)^3,$$

so that in three [[or more]] dimensions the random walk is transient.

The calculations are simple once we know that in one dimension $\mathbb{P}_0\{S_{2m} = 0\}$ is of order of magnitude $1/\sqrt{m}$. In a sense it is not very satisfactory to get this by using Stirling's formula and having huge exponentially large titans in the numerator and denominator fighting it out and killing each other off, leaving just a humble $\sqrt{m}$ standing in the denominator after the dust clears. In fact, it is easy to guess without any unnecessary violence or calculation that the order of magnitude is $1/\sqrt{m}$—note that the distribution of S_{2m}, having variance $2m$, is "spread out" over a range of order $\sqrt{m}$, so that the probabilities of points in that range should be of order $1/\sqrt{m}$. Another way to see the answer is to use a Normal approximation to the binomial distribution. We approximate the Binomial$(2m, 1/2)$ distribution by the Normal distribution $N(m, m/2)$, with the usual continuity correction:

$$\begin{aligned}\mathbb{P}\{\text{Binomial}(2m,1/2) = m\} &\sim \mathbb{P}\{m - 1/2 < N(m, m/2) < m + 1/2\}\\ &= \mathbb{P}\{-(1/2)\sqrt{2/m} < N(0,1) < (1/2)\sqrt{2/m}\}\\ &\sim \phi(0)\sqrt{2/m} = (1/\sqrt{2\pi})\sqrt{2/m} = 1/\sqrt{\pi m}.\end{aligned}$$

Although this calculation does not follow as a direct consequence of the usual Central Limit Theorem, it is an example of a "local Central Limit Theorem." □

(1.41) EXERCISE [THE OTHER 3-DIMENSIONAL RANDOM WALK]. *Consider a random walk on the 3-dimensional integer lattice; at each time the random walk moves with equal probability to one of the 6 nearest neighbors, adding or subtracting 1 in just one of the three coordinates. Show that this random walk is transient.*

Hint: You want to show that some series converges. An upper bound on the terms will be enough. How big is the largest probability in the Multinomial$(n; 1/3, 1/3, 1/3)$ *distribution?*

Here are a few additional problems about a simple symmetric random walk $\{S_n\}$ in one dimension starting from $S_0 = 0$ at time 0.

(1.42) EXERCISE. *Let a and b be integers with $a < 0 < b$. Defining the hitting times $\tau_c = \inf\{n \geq 0 : S_n = c\}$, show that the probability $\mathbb{P}\{\tau_b < \tau_a\}$ is given by $(0-a)/(b-a)$. Show that $\mathbb{P}\{\}$*

(1.43) EXERCISE. *Let $S_0, S_1, \ldots$ be a simple, symmetric random walk in one dimension as we have discussed, with $S_0 = 0$. Show that*

$$\mathbb{P}\{S_1 \neq 0, \ldots, S_{2n} \neq 0\} = \mathbb{P}\{S_{2n} = 0\}.$$

Now you can do a calculation that explains why the expected time to return to 0 is infinite.

(1.44) EXERCISE. *As in the previous exercise, consider a simple, symmetric random walk started out at 0. Letting $k \neq 0$ be any fixed state, show that the expected number of times the random walk visits state k before returning to state 0 is 1.*

We'll end this section with a discussion of the relationship between recurrence and the existence of a stationary distribution. The results will be useful in the next section.

(1.45) PROPOSITION. *Suppose a Markov chain has a stationary distribution π. If the state j is transient, then $\pi(j) = 0$.*

PROOF: Since π is stationary, we have $\pi P^n = \pi$ for all n, so that

$$\sum_i \pi(i) P^n(i,j) = \pi(j) \quad \text{for all } n. \tag{1.46}$$

However, since j is transient, Corollary (1.37) says that $\lim_{n\to\infty} P^n(i,j) = 0$ for all i. Thus, the left side of (1.46) approaches 0 as n approaches ∞, which implies that $\pi(j)$ must be 0. □

The last bit of reasoning about equation (1.46) may look a little strange, but in fact $\pi(i)P^n(i,j) = 0$ for all i and n. In light of what we now know, this is easy to see. Firstly, if i is transient, then $\pi(i) = 0$. Otherwise, if i is recurrent, then $P^n(i,j) = 0$ for all n, since if not, then j would be accessible from i, which would contradict the assumption that j is transient.

(1.47) COROLLARY. *If an irreducible Markov chain has a stationary distribution, then the chain is recurrent.*

PROOF: Being irreducible, the chain must be either recurrent or transient. However, if the chain were transient, then the previous Proposition would imply that $\pi(j) = 0$ for all j, which would contradict the assumption that π is a probability distribution, and so must sum to 1. □

The previous Corollary says that for an irreducible Markov chain, the existence of a stationary distribution implies recurrence. However, we know that the converse is not true. That is, there are irreducible, recurrent Markov chains that do not have stationary distributions. For example, we have seen that the simple symmetric random walk on the integers in one dimension is irreducible and recurrent but does not have a stationary distribution. This random walk is recurrent all right, but in a sense it is "just barely recurrent." That is, by recurrence we have $\mathbb{P}_0\{T_0 < \infty\} = 1$, for example, but we also have $\mathbb{E}_0(T_0) = \infty$. The name for this kind of recurrence is *null recurrence*: the state i is null recurrent if it is recurrent and $\mathbb{E}_i(T_i) = \infty$. Otherwise, a recurrent state is called *positive recurrent*: the state i is positive recurrent if $\mathbb{E}_i(T_i) < \infty$. A positive recurrent state i is not just barely recurrent, it is recurrent by a comfortable margin—when started at i, we have not only that T_i is finite almost surely, but also that T_i has finite expectation.

Positive recurrence is in a sense the right notion to relate to the existence of a stationary distibution. For now let me state just the facts, ma'am; these will be justified later. Positive recurrence is also a class property, so that if a chain is irreducible, the chain is either transient, null recurrent, or positive recurrent. It turns out that an irreducible chain has a stationary distribution if and only if it is positive recurrent. That is, strengthening "recurrence" to "positive recurrence" gives the converse to Corollary (1.47).

1.7 An aside on coupling

Coupling is a powerful technique in probability. It has a distinctly probabilistic flavor. That is, using the coupling idea entails thinking probabilistically, as opposed to simply applying analysis or algebra or some other area of mathematics. Many people like to prove assertions using coupling and feel happy when they have done so—a probabilisitic assertion deserves a probabilistic proof, and a good coupling proof can make obvious what might otherwise

be a mysterious statement. For example, we will prove the Basic Limit Theorem of Markov chains using coupling. As I have said before, we could do it using matrix theory, but the probabilist tends to find the coupling proof much more appealing, and I hope you do too.

It is a little hard to give a crisp definition of coupling, and different people vary in how they use the word and what they feel it applies to. Let's start by discussing a very simple example of coupling, and then say something about what the common ideas are.

(1.48) EXAMPLE [CONNECTIVITY OF A RANDOM GRAPH]. A graph is said to be *connected* if for each pair of distinct nodes i and j there is a path from i to j that consists of edges of the graph.

Consider a random graph on a given finite set of nodes, in which each pair of nodes is joined by an edge independently with probability p. We could simulate, or "construct," such a random graph as follows: for each pair of nodes $i < j$, generate a random number $U_{ij} \sim U[0,1]$, and join nodes i and j with an edge if $U_{ij} \leq p$. Here is a problem: show that the probability of the resulting graph being connected is nondecreasing in p. That is, for $p_1 < p_2$, we want to show that

$$\mathbb{P}_{p_1}\{\text{graph connected}\} \leq \mathbb{P}_{p_2}\{\text{graph connected}\}.$$

I would say that this is intuitively obvious, but we want to give an actual *proof*. Again, the example is just meant to illustrate the idea of coupling, not to give an example that can be solved only with coupling!

One way that one might approach this problem is to try to find an explicit expression for the probability of being connected as a function of p. Then one would hope to show that that function is increasing, perhaps by differentiating with respect to p and showing that the derivative is nonnegative.

That is conceptually a straightforward approach, but you may become discouraged at the first step—I don't think there is an obvious way of writing down the probability the graph is connected. Anyway, doesn't it seem somehow very inefficient, or at least "overkill," to have to give a precise expression for the desired probability if all one desires is to show the inituitively obvious monotonicity property? Wouldn't you hope to give an argument that somehow simply formalizes the intuition that we all have?

One nice way to show that probabilities are ordered is to show that the corresponding events are ordered: if $A \subseteq B$ then $\mathbb{P}A \leq \mathbb{P}B$. So let's make two events by making two random graphs G_1 and G_2, with each edge of G_1 having probability p_1 and each edge of G_2 having probability p_2. We could do that by using two sets of $U[0,1]$ random variables: $\{U_{ij}\}$ for G_1 and $\{V_{ij}\}$ for G_2. OK, so now we ask: is it true that

$$\{G_1 \text{ connected}\} \subseteq \{G_2 \text{ connected}\}? \tag{1.49}$$

The answer is no; indeed, the random graphs G_1 and G_2 are independent, so that clearly

$$\mathbb{P}\{G_1 \text{ connected}, G_2 \text{ not connected}\} = \mathbb{P}\{G_1 \text{ connected}\}\mathbb{P}\{G_2 \text{ not connected}\} > 0.$$

The problem is that we have used different, independent random numbers in constructing the graphs G_1 and G_2, so that, for example, it is perfectly possible to have simultaneously $U_{ij} \leq p_1$ and $V_{ij} > p_2$ for all $i < j$, in which the graph G_1 would be completely connected and the graph G_2 would be completely disconnected.

Here is a simple way to fix the argument: use the *same random numbers* in defining the two graphs. That is, draw the edge (i,j) in graph G_1 if $U_{ij} \leq p_1$ and the edge (i,j) in graph G_2 if $U_{ij} \leq p_2$. Now notice how the picture has changed: with the modified definitions it is obvious that, if an edge (i,j) is in the graph G_1, then that edge is also in G_2. From this, it is equally obvious that (1.49) now holds. This establishes the desired monotonicity of the probability of being connected. Perfectly obvious, isn't it? □

So, what characterizes a coupling argument? In our example, we wanted to establish a statement about two distributions: the distributions of random graphs with edge probabilities p_1 and p_2. To do this, we showed how to "construct" [[i.e., *simulate* using uniform random numbers!]] random objects having the desired distributions in such a way that the desired conclusion became obvious. The trick was to make appropriate use of the same uniform random variables in constructing the two objects. I think this is a general feature of coupling arguments: somewhere in there you will find the same set of random variables used to construct two different objects about which one wishes to make some probabilistic statement. The term "coupling" reflects the fact that the two objects are related in this way. □

(1.50) EXERCISE. *Consider a Markov Chain on the nonnegative integers $\mathbb{S} = \{0, 1, 2, \ldots\}$. Defining $P(i, i+1) = p_i$ and $P(i, i-1) = q_i$, assume that $p_i + q_i = 1$ for all $i \in \mathbb{S}$, and also $p_0 = 1$, $q_0 = 0$, and both p_i and q_i are positive for all $i \geq 1$. Use what you know about the simple, symmetric random walk to show that the given Markov chain is recurrent.*

1.8 Proof of the Basic Limit Theorem

The Basic Limit Theorem says that if an irreducible, aperiodic Markov chain has a stationary distribution π, then for each initial distribution π_0, as $n \to \infty$ we have $\pi_n(i) \to \pi(i)$ for all states i. Let me start by pointing something out, just in case the wording of the statement strikes you as a bit strange. Why does the statement read "... ***a*** stationary distribution"? For example, what if the chain has two stationary distributions? The answer is that this is impossible: the assumed conditions imply that a stationary distribution is in fact unique. In fact, once we prove the Basic Limit Theorem, we will know this to be the case. Clearly if the Basic Limit Theorem is true, an irreducible and aperodic Markov chain cannot have two different stationary distributions π and $\tilde{\pi}$, since obviously $\pi_n(i)$ cannot approach both $\pi(i)$ and $\tilde{\pi}(i)$ for all i.

An equivalent but conceptually useful reformulation is to define a distance between probability distributions, and then to show that as $n \to \infty$, the distance between the distribution π_n and the distribution π converges to 0. The notion of distance that we will use is called "total variation distance."

(1.51) DEFINITION. *Let λ and μ be two probability distributions on the set $\mathcal{S}$. Then the* ***total variation distance*** *$\|\lambda - \mu\|$ between λ and μ is defined by*

$$\|\lambda - \mu\| = \sup_{A \subset \mathcal{S}} [\lambda(A) - \mu(A)].$$

(1.52) EXERCISE. *Show that $\|\lambda - \mu\|$ may also be expressed in the alternative forms*

$$\|\lambda - \mu\| = \sup_{A \subset \mathcal{S}} |\lambda(A) - \mu(A)| = \frac{1}{2} \sum_{i \in \mathcal{S}} |\lambda(i) - \mu(i)| = 1 - \sum_{i \in \mathcal{S}} \min\{\lambda(i), \mu(i)\}.$$

Two probability distributions λ and μ assign probabilites to all possible events. The total variation distance between λ and μ is the largest possible discrepancy between the probabilities assigned by λ and μ to any event. For example, let π_7 denote the distribution of the ordering of a deck of cards after 7 shuffles, and let π denote the uniform distribution on all 52! permutations of the deck, which corresponds to the result of perfect shuffling (or "shuffling infinitely many times"). Suppose, for illustration, that the total variation distance $\|\pi_7 - \pi\|$ happens to be 0.17. This tells us that the probability of any event — for example, the probability of winning any specified card game — using a deck shuffled 7 times differs by at most 0.17 from the probability of the same event using a perfectly shuffled deck.

(1.53) EXERCISE. *Let π_0 and ρ_0 be probability mass functions on $\mathcal{S}$, and define $\pi_1 = \pi_0 P$ and $\rho_1 = \rho_0 P$, where P is a probability transition matrix. Show that $\|\pi_1 - \rho_1\| \leq \|\pi_0 - \rho_0\|$.*

To introduce the coupling method, let $Y_0, Y_1, \ldots$ be a Markov chain with the same probability transition matrix as $X_0, X_1, \ldots$, but let Y_0 have the distribution π; that is, we start the Y chain off in the initial distribution π instead of the initial distribution π_0 of the X chain. Note that $\{Y_n\}$ is a stationary Markov chain, and, in particular, that Y_n has the distribution π for all n. Further let the Y chain be independent of the X chain.

Roughly speaking, we want to show that for large n, the probabilistic behavior of X_n is close to that of Y_n. The next result says that we can do this by showing that for large n, the X and Y chains have met with high probability by time n. Define the *coupling time* T to be the first time at which X_n equals Y_n:

$$T = \inf\{n : X_n = Y_n\},$$

where of course we define $T = \infty$ if $X_n \neq Y_n$ for all n.

(1.54) LEMMA ["THE COUPLING INEQUALITY"]. *For all n we have*

$$\|\pi_n - \pi\| \leq \mathbb{P}\{T > n\}.$$

PROOF: Define the process $\{Y_n^*\}$ by

$$Y_n^* = \begin{cases} Y_n & \text{if } n < T \\ X_n & \text{if } n \geq T. \end{cases}$$

It is easy to see that $\{Y_n^*\}$ is a Markov chain, and it has the same probability transition matrix $P(i,j)$ as $\{X_n\}$ has. ⟦To understand this, start by thinking of the X chain as a frog carrying a table of random numbers jumping around in the state space. The frog uses his table of *iid* uniform random numbers to generate his path as we described earlier in the section about specifying and simulating Markov chains. He uses the first number in his table together with his initial distribution π_0 to determine X_0, and then reads down successive numbers in the table to determine the successive transitions on his path. The Y frog does the same sort of thing, except he uses his own, different table of uniform random numbers so he will be independent of the X frog, and he starts out with the initial distribution π instead of π_0. How about the Y^* frog? Is he also doing a Markov chain? Well, is he choosing his transitions using uniform random numbers like the other frogs? Yes, he is; the only difference is that he starts by using Y's table of random numbers (and hence he follows Y) until the coupling time T, after which he stops reading numbers from Y's table and switches to X's table. But big deal; he is still generating his path by using uniform random numbers in the way required to generate a Markov chain.⟧ The chain $\{Y_n^*\}$ is stationary: $Y_0^* \sim \pi$, since $Y_0^* = Y_0$ and $Y_0 \sim \pi$. Thus, $Y_n^* \sim \pi$ for all n. so that for $A \subseteq \mathcal{S}$ we have

$$\begin{aligned} \pi_n(A) - \pi(A) &= \mathbb{P}\{X_n \in A\} - \mathbb{P}\{Y_n^* \in A\} \\ &= \mathbb{P}\{X_n \in A, T \leq n\} + \mathbb{P}\{X_n \in A, T > n\} \\ &\quad -\mathbb{P}\{Y_n^* \in A, T \leq n\} - \mathbb{P}\{Y_n^* \in A, T > n\}. \end{aligned}$$

However, on the event $\{T \le n\}$, we have $Y_n^* = X_n$, so that the two events $\{X_n \in A, T \le n\}$ and $\{Y_n^* \in A, T \le n\}$ are the same, and hence they have the same probability. Therefore, the first and third terms in the last expression cancel, yielding

$$\pi_n(A) - \pi(A) = \mathbb{P}\{X_n \in A, T > n\} - \mathbb{P}\{Y_n^* \in A, T > n\}.$$

Since the last difference is obviously bounded by $\mathbb{P}\{T > n\}$, we are done. □

Note the significance of the coupling inequality: it reduces the problem of showing that $\|\pi_n - \pi\| \to 0$ to that of showing that $\mathbb{P}\{T > n\} \to 0$, or equivalently, that $\mathbb{P}\{T < \infty\} = 1$. To do this, we consider the "bivariate chain" $\{Z_n = (X_n, Y_n) : n \ge 0\}$. A bit of thought confirms that $Z_0, Z_1, \ldots$ is a Markov chain on the state space $\mathcal{S} \times \mathcal{S}$. Since the X and Y chains are independent, the probability transition matrix P_Z of the Z chain can be written as

$$P_Z(i_x i_y, j_x j_y) = P(i_x, j_x)P(i_y, j_y).$$

It is easy to check that the Z chain has stationary distribution

$$\pi_Z(i_x i_y) = \pi(i_x)\pi(i_y).$$

Watch closely now; we're about to make an important reduction of the problem. Recall that we want to show that $\mathbb{P}\{T < \infty\} = 1$. Stated in terms of the Z chain, we want to show that with probability one, the Z chain hits the "diagonal" $\{(j, j) : j \in \mathcal{S}\}$ in $\mathcal{S} \times \mathcal{S}$ in finite time. To do this, it is sufficient to show that the Z chain is irreducible and recurrent ⟦why?⟧. However, since we know that the Z chain has a stationary distribution, by Corollary (1.47), to prove the Basic Limit Theorem, it suffices to show that the Z chain is irreducible.

This is, strangely†, the hard part. This is where the aperiodicity assumption comes in. For example, consider a Markov chain $\{X_n\}$ having the "type A frog" transition matrix $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ started out in the condition $X_0 = 0$. Then the stationary chain $\{Y_n\}$ starts out in the uniform distribution: probability 1/2 on each state 0,1. The bivariate chain $\{(X_n, Y_n)\}$ is not irreducible: for example, from the state $(0, 0)$, we clearly cannot reach the state $(0, 1)$. And this ruins everything. For example, if $Y_0 = 1$, which happens with probability 1/2, *the X and Y chains can never meet,* so that $T = \infty$. Thus, $\mathbb{P}\{T < \infty\} < 1$.

A little number-theoretic result will help us establish irreducibility of the Z chain.

(1.55) LEMMA. *Suppose A is a set of positive integers that is closed under addition and has greatest common divisor (gcd) one. Then there exists an integer N such that $n \in A$ for all $n \ge N$.*

PROOF: First we claim that A contains at least one pair of consecutive integers. To see this, suppose to the contrary that the minimal "spacing" between successive elements of A is $s > 1$. That is, any two distinct elements of A differ by at least s, and there exists an integer n_1 such that both $n_1 \in A$ and $n_1 + s \in A$. Let $m \in A$ be such that s does not

†Or maybe not so strangely, in view of Example (1.32).

divide m; we know that such an m exists because $\gcd(A) = 1$. Write $m = qs + r$, where $0 < r < s$. Now observe that, by the closure under addition assumption, the two numbers $a_1 = (q+1)(n_1+s)$ and $a_2 = (q+1)n_1 + m$ are both in A. However, $a_1 - a_2 = s - r \in (0, s)$, which contradicts the definition of s. This proves the claim.

Thus, A contains two consecutive integers, say, c and $c+1$. Now we will finish the proof by showing that $n \in A$ for all $n \geq c^2$. If $c = 0$ this is trivially true, so assume that $c > 0$. We have, by closure under addition,

$$\begin{aligned} c^2 &= (c)(c) \in A \\ c^2 + 1 &= (c-1)c + (c+1) \in A \\ &\vdots \\ c^2 + c - 1 &= c + (c-1)(c+1) \in A. \end{aligned}$$

Thus, $\{c^2, c^2+1, \ldots, c^2+c-1\}$, a set of c consecutive integers, is a subset of A. Now we can add c to all of these numbers to show that the next set $\{c^2+c, c^2+c+1, \ldots, c^2+2c-1\}$ of c integers is also a subset of A. Repeating this argument, clearly all integers c^2 or above are in A. □

Let $i \in \mathcal{S}$, and retain the assumption that the chain is aperiodic. Then since the set $\{n : P^n(i,i) > 0\}$ is clearly closed under addition, and, by the aperiodicity assumption, has greatest common divisor 1, the previous lemma applies to give that $P^n(i,i) > 0$ for all sufficiently large n. From this, for any $i, j \in \mathcal{S}$, since irreducibility implies that $P^m(i,j) > 0$ for some m, it follows that $P^n(i,j) > 0$ for all sufficiently large n.

Now we complete the proof of the Basic Limit Theorem by showing that the chain $\{Z_n\}$ is irreducible. Let $i_x, i_y, j_x, j_y \in \mathcal{S}$. It is sufficient to show, in the bivariate chain $\{Z_n\}$, that $(j_x j_y)$ is accessible from $(i_x i_y)$. To do this, it is sufficient to show that $P_Z^n(i_x i_y, j_x j_y) > 0$ for some n. However, by the assumed independence of $\{X_n\}$ and $\{Y_n\}$,

$$P_Z^n(i_x i_y, j_x j_y) = P^n(i_x, j_x) P^n(i_y, j_y),$$

which, by the previous paragraph, is positive for all sufficiently large n. Of course, this implies the desired result, and we are done.

(1.56) EXERCISE. *[[A little practice with the coupling idea]]*

(i) *Consider a Markov chain $\{X_n\}$ having probability transition matrix*

$$P = \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/4 & 1/2 & 1/4 \\ 1/4 & 1/4 & 1/2 \end{pmatrix}.$$

Note that $\{X_n\}$ has stationary distribution $\pi = (1/3, 1/3, 1/3)$. Using the sort of coupling we did in the proof of the Basic Limit Theorem, show that, no matter what the initial distribution π_0 of X_0 is, we have

$$\|\pi_n - \pi\| \leq \frac{2}{3}\left(\frac{11}{16}\right)^n$$

for all n.

(ii) *Do you think the bound you just derived is a good one? In particular, is 11/16 the smallest we can get? What is the best we could do?*

(iii) *Can you use a more "aggressive" coupling to get a better bound? [[What do I mean? The coupling we used in the proof of the Basic Limit Theorem was not very aggressive, in that it let the two chains evolve independently until they happened to meet, and only then started to use the same uniform random numbers to generate the paths. No attempt was made to get the chains together as fast as possible. A more aggressive coupling would somehow make use of some random numbers in common to both chains in generating their paths right from the beginning.]]*

1.9 A SLLN for Markov chains

The usual Strong Law of Large Numbers for independent and identically distributed (*iid*) random variables says that if $X_1, X_2, \ldots$ are *iid* with mean μ, then the average $(1/n)\sum_{t=1}^{n} X_t$ converges to μ with probability 1 as $n \to \infty$.

> Some fine print: It is possible to have $\mu = +\infty$, and the SLLN still holds. For example, supposing that the random variables X_t take their values in the set of nonnegative integers $\{0, 1, 2, \ldots\}$, the mean is defined to be $\mu = \sum_{k=0}^{\infty} k\mathbb{P}\{X_0 = k\}$. This sum could diverge, in which case we define μ to be $+\infty$, and we have $(1/n)\sum_{t=1}^{n} X_t \to \infty$ with probability 1.

For example, if $X_0, X_1, \ldots$ are *iid* with values in the set $\mathbb{S}$, then the SLLN tells us that

$$(1/n)\sum_{t=1}^{n} I\{X_t = i\} \to \mathbb{P}\{X_0 = i\}$$

with probability 1 as $n \to \infty$. That is, the fraction of times that the *iid* process takes the value i in the first n observations converges to $\mathbb{P}\{X_0 = i\}$, the probability that any given observation is i.

We will do a generalization of this result for Markov chains. This law of large numbers will tell us that the fraction of times that a Markov chains occupies state i converges to a limit.

It is possible to view this result as a consequence of a more general and rather advanced *ergodic theorem* (see, for example, Durrett's *Probability: Theory and Examples*). However, I do not want to assume prior knowledge of ergodic theory. Also, the result for Markov chains is quite simple to derive as a consequence of the ordinary law of large numbers for *iid* random variables. Although the successive states of a Markov chain are not independent, of course, we have seen that certain features of a Markov chain are independent of each other. Here we will use the idea that the path of the chain consists of a succession of independent "cycles," the segments of the path between successive visits to a recurrent state. This independence makes the treatment of Markov chains simpler than the general treatment of stationary processes, and it allows us to apply the law of large numbers that we already know.

(1.57) THEOREM. *Let $X_0, X_1, \ldots$ be a Markov chain starting in the state $X_0 = i$, and suppose that the state i communicates with another state j. The limiting fraction of time that the chain spends in state j is $1/\mathbb{E}_j T_j$. That is,*

$$\mathbb{P}_i \left\{ \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} I\{X_t = j\} = \frac{1}{\mathbb{E}_j T_j} \right\} = 1.$$

PROOF: The result is easy if the state j is transient, since in that case $\mathbb{E}_j T_j = \infty$ and (with probability 1) the chain visits j only finitely many times, so that

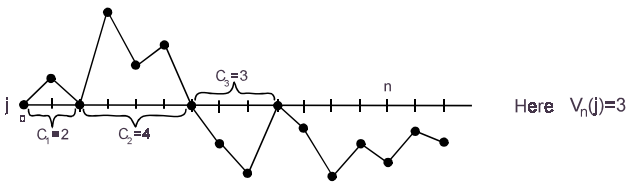
$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} I\{X_t = j\} = 0 = \frac{1}{\mathbb{E}_j T_j}$$

with probability 1. So we assume that j is recurrent. We will also begin by proving the result in the case $i = j$; the general case will be an easy consequence of this special case. Again we will think of the Markov chain path as a succession of *cycles*, where a cycle is a segment of the path that lies between successive visits to j. The cycle lengths $C_1, C_2, \ldots$ are *iid* and distributed as T_j; here we have already made use of the assumption that we are starting at the state $X_0 = j$. Define $S_k = C_1 + \cdots + C_k$ and let $V_n(j)$ denote the number of visits to state j made by $X_1, \ldots, X_n$, that is,

$$V_n(j) = \sum_{t=1}^{n} \{X_t = j\}.$$

A bit of thought [[see also the picture below]] shows that $V_n(j)$ is also the number of cycles completed up to time n, that is,

$$V_n(j) = \max\{k : S_k \le n\}.$$



To ease the notation, let V_n denote $V_n(j)$. Notice that

$$S_{V_n} \le n < S_{V_n+1},$$

and divide by V_n to obtain

$$\frac{S_{V_n}}{V_n} \leq \frac{n}{V_n} < \frac{S_{V_n+1}}{V_n}.$$

Since j is recurrent, $V_n \to \infty$ with probability one as $n \to \infty$. Thus, by the ordinary Strong Law of Large Numbers for *iid* random variables, we have both

$$\frac{S_{V_n}}{V_n} \to \mathbb{E}_j(T_j)$$

and

$$\frac{S_{V_n+1}}{V_n} = \left(\frac{S_{V_n+1}}{V_n+1}\right)\left(\frac{V_n+1}{V_n}\right) \to \mathbb{E}_j(T_j) \times 1 = \mathbb{E}_j(T_j)$$

with probability one. Note that the last two displays hold whether $\mathbb{E}_j(T_j)$ is finite or infinite. Thus, $n/V_n \to \mathbb{E}_j(T_j)$ with probability one, so that

$$\frac{V_n}{n} \to \frac{1}{\mathbb{E}_j T_j}$$

with probability one, which is what we wanted to show.

Next, to treat the general case where i may be different from j, note that $P_i\{T_j < \infty\} = 1$ by Theorem 1.35. Thus, with probability one, a path starting from i behaves as follows. It starts by going from i to j in some finite number T_j of steps, and then proceeds on from state j in such a way that the long run fraction of time that $X_t = j$ for $t \geq T_j$ approaches $1/\mathbb{E}_j(T_j)$. But clearly the long run fraction of time the chain is at j is not affected by the behavior of the chain on the finite segment $X_0, \ldots, X_{T_j-1}$. So with probability one, the long run fraction of time that $X_n = j$ for $n \geq 0$ must approach $1/\mathbb{E}_j(T_j)$. □

The following result follows directly from Theorem (1.57) by the Bounded Convergence Theorem from the Appendix. ⟦That is, we are using the following fact: if $Z_n \to c$ with probability one as $n \to \infty$ and the random variables Z_n all take values in the same bounded interval, then we also have $\mathbb{E}(Z_n) \to c$. To apply this in our situation, note that we have

$$Z_n := \frac{1}{n}\sum_{t=1}^{n} I\{X_t = j\} \to \frac{1}{\mathbb{E}_j T_j}$$

with probability one as $n \to \infty$, and also each Z_n lies in the interval [0,1]. Finally, use the fact that the expectation of an indicator random variable is just the probability of the corresponding event.⟧

(1.58) COROLLARY. *For an irreducible Markov chain, we have*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} P^t(i,j) = \frac{1}{\mathbb{E}_j(T_j)}$$

for all states i and j.

There's something suggestive here. Consider for the moment an irreducible, aperiodic Markov chain having a stationary distribution π. From the Basic Limit Theorem, we know that, $P^n(i,j) \to \pi(j)$ as $n \to \infty$. However, it is simple fact that if a sequence of numbers converges to a limit, then the sequence of "Cesaro averages" converges to the same limit; that is, if $a_t \to a$ as $t \to \infty$, then $(1/n)\sum_{t=1}^n a_t \to a$ as $n \to \infty$. Thus, the Cesaro averages of $P^n(i,j)$ must converge to $\pi(j)$. However, the previous Corollary shows that the Cesaro averages converge to $1/\mathbb{E}_j(T_j)$. Thus, it follows that

$$\pi(j) = \frac{1}{\mathbb{E}_j(T_j)}.$$

It turns out that the aperiodicity assumption is not needed for this last conclusion; we'll see this in the next result. Incidentally, we could have proved this result much earlier; for example we don't need the Basic Limit Theorem in the development.

(1.59) THEOREM. *An irreducible, positive recurrent Markov chain has a unique stationary distribution π given by*

$$\pi(j) = \frac{1}{\mathbb{E}_j(T_j)}.$$

PROOF: For the uniqueness, let π be a stationary distribution. We start with the relation

$$\sum_i \pi(i)P^t(i,j) = \pi(j),$$

which holds for all t. Averaging this over values of t from 1 to n gives

$$\sum_i \pi(i)\frac{1}{n}\sum_{t=1}^n P^t(i,j) = \pi(j).$$

By Corollary 1.58 ⟦and the Dominated Convergence Theorem⟧, the left side of the last equation approaches

$$\sum_i \pi(i)\frac{1}{\mathbb{E}_j(T_j)} = \frac{1}{\mathbb{E}_j(T_j)}$$

as $n \to \infty$. Thus, $\pi(j) = 1/\mathbb{E}_j(T_j)$, which establishes the uniqueness assertion.

We begin the proof of existence by doing the proof in the special case where the state space is finite. The proof is simpler here than in the general case, which involves some distracting technicalities.

So assume for the moment that the state space is finite. We begin again with Corollary 1.58, which says that

$$\frac{1}{n}\sum_{t=1}^n P^t(i,j) \to \frac{1}{\mathbb{E}_j(T_j)}. \tag{1.60}$$

However, the sum over all j of the left side of (1.60) is 1, for all n. Therefore,

$$\sum_j \frac{1}{\mathbb{E}_j(T_j)} = 1.$$

That's good, since we want our claimed stationary distribution to be a probability distribution.

Next we write out the matrix equation $P^t P = P^{t+1}$ as follows:

$$\sum_k P^t(i,k)P(k,j) = P^{t+1}(i,j). \tag{1.61}$$

Averaging this over $t = 1, \ldots, n$ gives

$$\sum_k \left[\frac{1}{n}\sum_{t=1}^{n} P^t(i,k)\right] P(k,j) = \frac{1}{n}\sum_{t=1}^{n} P^{t+1}(i,j).$$

Taking the limit as $n \to \infty$ of the last equation and using (1.60) again gives

$$\sum_k \left(\frac{1}{\mathbb{E}_k T_k}\right) P(k,j) = \frac{1}{\mathbb{E}_j T_j}.$$

Thus, our claimed stationary distribution is indeed stationary.

Finally, let's see how to handle the infinite state space case. Let $A \subset \mathcal{S}$ be a finite subset of the state space. Summing (1.60) over $j \in A$ gives the inequality

$$\sum_{j\in A} \frac{1}{\mathbb{E}_j(T_j)} \leq 1.$$

Therefore, since this is true for all subsets A, we get

$$\sum_{j\in \mathcal{S}} \frac{1}{\mathbb{E}_j(T_j)} =: C \leq 1.$$

By the assumption of positive recurrence, we have $C > 0$; in a moment we'll see that $C = 1$. The same sort of treatment of (1.61) [[i.e., sum over $k \in A$, average over $t = 1, \ldots, n$, let $n \to \infty$, and then take supremum over subsets A of $\mathcal{S}$]] gives the inequality

$$\sum_k \left(\frac{1}{\mathbb{E}_k T_k}\right) P(k,j) \leq \frac{1}{\mathbb{E}_j T_j}. \tag{1.62}$$

However, the sum over all j of the left side of (1.62) is

$$\sum_k \left(\frac{1}{\mathbb{E}_k T_k}\right) \sum_j P(k,j) = \sum_k \left(\frac{1}{\mathbb{E}_k T_k}\right),$$

which is the same as the sum of the right side of (1.62). Thus, the left and right sides of (1.62) must be the same for all j. From this we may conclude that the distribution

$$\tilde{\pi}(j) = \frac{1}{C}\left(\frac{1}{\mathbb{E}_j(T_j)}\right)$$

is stationary, so that, in particular, we know that our chain does have a stationary distribution. Thus, by the uniqueness assertion we proved above, we must have $C = 1$, and we are done. □

(1.63) EXERCISE. *Consider a knight sitting on the lower left corner square of an ordinary 8×8 chess board. The knight has residual frog-like tendencies, left over from an old spell an older witch cast upon him. So he performs a random walk on the chess board, at each time choosing a random move uniformly distributed over the set of his possible knight moves. What is the expected time until he first returns to the lower left corner square?*

(1.64) EXERCISE. *Recall the definition of positive recurrence on page 1-22. Show that positive recurrence is a class property.*

(1.65) EXERCISE. *Suppose a Markov chain has a stationary distribution π and the state j is null recurrent. Show that $\pi(j) = 0$.*

(1.66) EXERCISE [BIRTH-COLLAPSE CHAIN]. *Consider a Markov chain on $\mathcal{S} = \{0, 1, 2, \ldots\}$ having $P(i, i+1) = p_i$, $P(i, 0) = 1 - p_i$ for all i, with $p_0 = 1$ and $0 < p_i < 1$ for all $i > 0$. Show that*

(i) *The chain is recurrent if and only if $\lim_{n\to\infty} \prod_{i=1}^{n} p_i = 0$. [[This, in turn, is equivalent to the condition $\sum_{i=1}^{\infty}(1 - p_i) = \infty$. (This was just for interest; not a problem or a hint.)]]*

(ii) *The chain is positive recurrent if and only if $\sum_{n=1}^{\infty} \prod_{i=1}^{n} p_i < \infty$.*

(iii) *What is the stationary distribution if $p_i = 1/(i+1)$?*

1.10 General state space Markov chains

So far we have been discussing Markov chains with finite or countably infinite state spaces. But many applications are most naturally modeled as processes moving on more general state spaces, such as the real line or higher dimensional Euclidean spaces.

WARNING: This section may be rather long and tiring. It should probably be revised and streamlined... Suggestions welcome.

(1.67) EXAMPLE. Another standard use of the term "random walk" is for a sequence of partial sums of iid random variables. For example, we might have $Z_1, Z_2, \ldots$ independent and distributed according to the normal distribution $N(\mu, 1)$ with mean μ and variance 1,

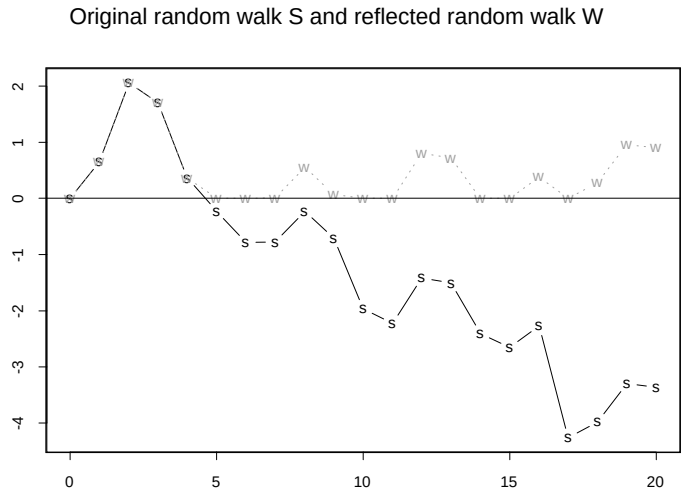
and define the *random walk* $X_0, X_1, \ldots$ by $X_n = Z_1 + \cdots + Z_n$ for $n \geq 0$. In contrast with the simple symmetric random walk, which moves around on the integers, such a normal random walk has probability 0 of being in any given countable set of numbers at any positive time. □

(1.68) EXAMPLE [AUTOREGRESSIVE PROCESS]. Autoregressive processes are the bread and butter of time series analysis. Here is a simple example. Let X_0 have a Normal distribution $N(\mu_0, \sigma_0^2)$, and define $X_1, X_2, \ldots$ recursively by $X_t = \theta X_{t-1} + Z_t$, where $Z_1, Z_2, \ldots$ are iid $N(0, \tau^2)$. Then $\{X_t\}$ is an example of an autoregressive process of order 1. □

(1.69) EXAMPLE [REFLECTED RANDOM WALK]. Let $X_1, X_2, \ldots$ be *iid*, and define the process $\{W_t\}$ by the recursion

$$W_t = \max\{0, W_{t-1} + X_t\} \quad \text{for } t > 0.$$

and $W_0 = 0$, say. Then $\{W_t\}$ is called a *reflected random walk*. The W process makes *iid* increments like a random walk, *except* when taking such an increment would cause the process to become negative, in which case the process takes the value 0. Reflected random walks arise in diverse contexts, including queueing theory and statistical procedures for quickly detecting a change in a probability distribution. As an example, if the random variables $X_1, X_2, \ldots$ are *iid* with distribution $N(\mu, 1)$, with the "drift" $\mu < 0$, then the reflected random walk keeps trying to drift downward and repeatedly bumps against the reflecting barrier at 0. An example with $\mu = -0.3$ is shown in the figure.



Notice a qualitative difference between this process and the previous examples: here we have an *atom*, in the sense that there is a state (0, here) that is hit with positive probability. □

A Markov chain $\{X_0, X_1, \ldots\}$ is determined by a state space $\mathcal{S}$, an initial distribution π_0, and a probability transition rule. The state space is a set, and the initial distribution is a probability measure on that set. For each $x \in \mathcal{S}$, the probability transition rule, or "transition kernel," specifies a probability measure on $\mathcal{S}$. That is, the transition kernel P of the chain gives conditional probabilities like

$$P(x, A) = \mathbb{P}\{X_{t+1} \in A \mid X_t = x\}.$$

Letting π_t denote the distribution of X_t, we have $\pi_{t+1} = \pi_t P$, that is,

$$\pi_{t+1}(A) = \int \pi_t(dx) P(x, A)$$

As you might suspect by now, much of the theory we have developed for countable state spaces extends to more general state spaces, with sums replaced by integrals.

A *stationary distribution* π is a probability distribution on $\mathcal{S}$ that satisfies the equation

$$\int \pi(dx) P(x, A) = \pi(A)$$

for all $A \subset \mathcal{S}$.

(1.70) Example [Autoregressive process, continued]. Continuing with Example (1.68), suppose $-1 < \theta < 1$. Sensibly suspecting the family of Normal distributions as the plausible candidates for a stationary distribution here, let us try out the distribution $\pi = N(\mu, \sigma^2)$ and see what the values of μ and σ have to be. Assuming X_{t-1} and X_t are distributed according to π and noting that Z_t is independent of X_{t-1}, by equating the means and variances of the left and right side of $X_t = \theta X_{t-1} + Z_t$ we obtain the equations $\mu = \theta\mu$ and $\sigma^2 = \theta^2\sigma^2 + \tau^2$, which imply $\mu = 0$ and $\sigma^2 = \tau^2/(1-\theta^2)$. Denoting the distribution at time t by $\pi_t = N(\mu_t, \sigma_t^2)$, we ask: does π_t approach π as $t \to \infty$? Let's compute μ_t and σ_t explicitly. Applying the relations $\mu_t = \theta\mu_{t-1}$ and $\sigma_t^2 = \theta^2\sigma_{t-1}^2 + \tau^2$ to $t = 1, 2, \ldots$ gives

$$\begin{aligned}
\mu_1 &= \theta\mu_0, \quad & \sigma_1^2 &= \theta^2\sigma_0^2 + \tau^2, \\
\mu_2 &= \theta^2\mu_0, \quad & \sigma_2^2 &= \theta^4\sigma_0^2 + \theta^2\tau^2 + \tau^2, \\
\mu_3 &= \theta^3\mu_0, \quad & \sigma_3^2 &= \theta^6\sigma_0^2 + \theta^4\tau^2 + \theta^2\tau^2 + \tau^2, \\
&\vdots & & \\
\mu_t &= \theta^t\mu_0, \quad & \sigma_t^2 &= \theta^{2t}\sigma_0^2 + (\theta^{2t-2} + \theta^{2t-4} + \cdots + \theta^2 + 1)\tau^2, \\
&\vdots & &
\end{aligned}$$

Thus, $\mu_t \to 0$ and $\sigma_t^2 \to \tau^2 \sum_{k=0}^{\infty} \theta^{2k} = \tau^2/(1-\theta^2)$, and we have established convergence to the stationary distribution $N(0, \tau^2/(1-\theta^2))$. So here is a continuous-state-space Markov

chain for which we have found a stationary distribution and established convergence to stationarity. □

The last example was nice and easy, but we have shamelessly exploited the special features of this problem. In particular, the Normality assumptions allowed us to do explicit computations of the distributions π_t and π. However, what happens, for example, if the random variables $\{Z_t\}$ are not Normally distributed? Presumably under some mild conditions we will still have convergence to the stationary distribution, whatever it is, but our simple calculations go out the window. Is there a general theorem we can appeal to, analogous to the Basic Limit Theorem we got in the discrete space case?

(1.71) EXAMPLE [MARKOV SAMPLING]. We have seen this idea before in discrete state spaces; it works more generally also. If we want to simulate a sample from a given probability distribution π on a set $\mathcal{S}$, the Basic Limit Theorem will tell us that we can do this approximately by running a Markov chain having state space $\mathcal{S}$ and stationary distribution π. There are a number of popular methods for manufacturing a Markov chain having a given desired distribution as its stationary distribution, such as the Metropolis method and the Gibbs sampler.

As discussed earlier, the Gibbs sampler proceeds by simulating from conditional distributions that are, one hopes, simpler to simulate than the original distribution. For example, suppose we wish to simulate from a given probability density function f on $\mathbb{R}^2$, which is an uncountable set, not discrete. For purposes of this discussion let (X, Y) denote a pair of random variables having joint density f. We would like to simulate such a pair of random variables, at least approximately. Given that we are now (time t) at the state $(X_t, Y_t) = (x, y)$, we could generate the next state (X_{t+1}, Y_{t+1}) as follows. Flip a coin. If Heads, let $X_{t+1} = X_t = x$, and draw Y_{t+1} from the conditional distribution of Y given $X = x$. If Tails, let $Y_{t+1} = Y_t = y$, and draw X_{t+1} from the conditional distribution of X given $Y = y$. The sequence $\{(X_t, Y_t) : t = 0, 1, \ldots\}$ is a Markov chain having stationary density f.

What we would like here is a general Basic Limit Theorem that would allow us to prove that the Gibbs sampler Markov chain converges in distribution to its stationary distribution. □

1.10.1 Chains with an atom

Do you remember our proof of the Basic Limit Theorem in the discrete case? We used the coupling idea: run two independent copies of the chain until they *couple*, that is, until they hit the same state at some time T. The coupling inequality $\|\pi_t - \pi\| \leq \mathbb{P}\{T > t\}$ reduced the problem of showing that $\|\pi_t - \pi\| \to 0$ to the problem of showing that $\mathbb{P}\{T < \infty\} = 1$. In other words, we reduced the problem to showing that with probability 1, the two chains eventually must couple. However, in typical examples in general state spaces, each

individual state is hit with probability 0, and independent copies of the chain will never couple. An *atom* is a state that is hit with positive probability. If a Markov chain has an atom, then we can hope to carry through the same sort of coupling argument as we used in the descrete case. In this section we develop a basic limit theorem for chains having an atom.

(1.72) DEFINITION. *An **accessible atom** α is a state that is hit with positive probability starting from each state, that is, $\sum_{t=0}^{\infty} \mathbb{P}_x\{X_t = \alpha\} > 0$ for all $x \in \mathcal{S}$.*

(1.73) EXAMPLE. In Example (1.69), the state 0 is an accessible atom. □

Our goal in this section is a Basic Limit Theorem for chains that have atoms. Although it is natural to think that most chains of interest do not have atoms, so that the theory developed in this section would not often apply, we will see in the next section how a surprisingly large class of chains may be viewed as chains with an atom.

(1.74) PROPOSITION. *Suppose a chain with an accessible atom α has a stationary distribution π. Then $\pi\{\alpha\} > 0$ and α is recurrent.*

PROOF: Since α is accessible, it follows that for each state x there is a t such that $P^t(x, \{\alpha\}) > 0$. That is, defining $G_t = \{x : P^t(x, \{\alpha\}) > 0\}$, we have $\bigcup G_t = \mathcal{S}$. So there is an n such that $\pi(G_n) > 0$, which gives

$$\pi\{\alpha\} = \int \pi(dx) P^n(x, \{\alpha\}) \geq \int_{G_n} \pi(dx) P^n(x, \{\alpha\}) > 0.$$

⟦The integral of a positive function over a set of positive measure is positive.⟧ The proof that α is recurrent is like what we did before for countable state spaces. Since $\mathbb{P}_\pi\{X_t = \alpha\} = \pi\{\alpha\} > 0$ for all t, defining $N_\alpha = \sum_{t=0}^{\infty} I\{X_t = \alpha\}$, we get $\mathbb{E}_\pi(N_\alpha) = \infty$. But $\mathbb{E}_\alpha(N_\alpha) \geq \mathbb{E}_x(N_\alpha)$ for all states x; recall that starting from α we get to count at least one visit to α for sure! So, averaging over π, we get $\mathbb{E}_\alpha(N_\alpha) \geq \int \pi(dx) \mathbb{E}_x(N_\alpha) = \mathbb{E}_\pi(N_\alpha)$, so that $\mathbb{E}_\alpha(N_\alpha) = \infty$. This implies the recurrence of α, by the geometric trials argument from before. □

(1.75) PROPOSITION. *Suppose the chain $\{X_t\}$ has an accessible atom α and a stationary distribution π. Let B be a set that is not accessible from α, that is, $\mathbb{P}_\alpha\{T_B < \infty\} = 0$. Then $\pi(B) = 0$.*

PROOF: Define

$$B_{\delta,n} = \{x \in B : \mathbb{P}_x\{T_\alpha \leq n\} \geq \delta\}.$$

By the assumption that α is an accessible atom, $\bigcup_{m,n} B_{1/m,n} = B$. Thus, we will be done if we show that $\pi(B_{\delta,n}) = 0$ for each n and each $\delta > 0$. So consider a fixed n and $\delta > 0$.

Starting from any $x \in B_{\delta,n}$, with probability at least δ, the chain goes to α within n steps, and then never returns to $B_{\delta,n}$. [[The last statement about not returning to $B_{\delta,n}$ follows by definition of B and the fact that $B_{\delta,n} \subseteq B$.]] So each time we enter $B_{\delta,n}$, there is probability at least δ that within n steps we leave $B_{\delta,n}$ forever. Defining $N = \sum_{t=0}^{\infty} I\{X_t \in B_{\delta,n}\}$ to be the total number of visits to the set $B_{\delta,n}$, a bit of thought shows that $\mathbb{E}_y N \le n/\delta$ for each $y \in \mathcal{S}$. [[Here is one way to see this. Look at the total number $N_0 = \sum_{r=0}^{\infty} I\{X_{rn} \in B_{\delta,n}\}$ of visits to $B_{\delta,n}$ at times $0, n, 2n, \ldots$. Then $\mathbb{P}_y\{N_0 > 1\} \le 1-\delta$, $\mathbb{P}_y\{N_0 > 2\} \le (1-\delta)^2$, and so on. So

$$\mathbb{E}_y N_0 = \sum_{r=0}^{\infty} \mathbb{P}\{N_0 > r\} \le \sum_{r=0}^{\infty} (1-\delta)^r = 1/\delta.$$

Similarly, for each $0 < k < n$, the number $N_k = \sum_{r=0}^{\infty} I\{X_{k+rn} \in B_{\delta,n}\}$ of visits at times $k, k+n, k+2n, \ldots$ satisfies $\mathbb{E}_y N_k \le 1/\delta$. Thus, $N = N_0 + N_1 + \cdots + N_{n-1}$ has expected value at most n/δ, starting from any state y.]] So $\mathbb{E}_\pi N \le n/\delta$. This implies that $\pi(B_{\delta,n}) = 0$: if $\pi(B_{\delta,n})$ were positive, then clearly $\mathbb{E}_\pi N$ would be infinite, which we have just shown is not the case. □

The previous result implies that a stationary chain with an accessible atom α will not enter a set of states that is not accessible from α.

(1.76) PROPOSITION. *Suppose a Markov chain has an accessible atom α and a stationary distribution π. Then $\mathbb{P}_\pi\{T_\alpha < \infty\} = 1$.*

PROOF: Let $B = \{x : \mathbb{P}_x\{T_\alpha = \infty\} > 0\}$; these are the states from which it is possible to go forever without hitting α. We want to show that $\pi(B) = 0$. Since α is recurrent, if the chain starts from state α, then with probability 1 it will return to α infinitely many times. Therefore, the set B cannot be accessible from α, for if it were, there would be positive probability, starting from α, of eventually entering the set B and then never returning to α. Thus, by the previous proposition, $\pi(B) = 0$. □

(1.77) DEFINITION. *Let μ and ν be two probability measures on a set $\mathcal{S}$. We say that μ is **absolutely continuous** with respect to ν if $\mu(A) = 0$ for all $A \subseteq \mathcal{S}$ such that $\nu(A) = 0$, that is, each set having probability 0 under ν also has probability 0 under μ.*

(1.78) THEOREM. *Suppose a chain $\{X_t\}$ with transition kernel P and an aperiodic, accessible atom α has a stationary distribution π. Let π_t denote the distribution of X_t and start the chain in any distribution π_0 that is absolutely continuous with respect to π. Then $\|\pi_t - \pi\| \to 0$ as $t \to \infty$.*

PROOF: We use the coupling technique from before; much of the reasoning remains the same, so I'll just give a sketch. Again, we run two independent copies of the chain, $\{X_t\}$ and $\{X_t^*\}$, starting in the initial distributions π_0 and π, respectively. We want to show that

with probability 1 the two chains eventually couple; in fact we claim that they eventually visit the state α at the same time. By using the aperiodicity assumption together with the number-theoretic lemma from before, we see that the bivariate chain $\{(X_t, X_t^*) : t = 0, 1, \ldots\}$ has an accessible atom (α, α). The bivariate chain has a stationary distribution: the obvious product distribution $\pi \times \pi$. So by Proposition 1.76, if the bivariate chain were started out in its stationary distribution $\pi \times \pi$, it would eventually hit its atom (α, α) with probability 1. That is, letting A denote the set of pairs of states (x, y) such that $\mathbb{P}_{(x,y)}\{T_{(\alpha,\alpha)} < \infty\} = 1$, we have $(\pi \times \pi)(A) = 1$. From this, the absolute continuity of π_0 with respect to π implies that $(\pi_0 \times \pi)(A) = 1$ [[observe that $(\pi \times \pi)(A^c) = 0$ implies $(\pi_0 \times \pi)(A) = 0$]]. Thus, $\mathbb{P}_{\pi_0 \times \pi}\{T_{(\alpha,\alpha)} < \infty\} = 1$, as claimed. □

(1.79) EXERCISE. *Do we really need the hypothesis about the absolute continuity of π_0? Here is an example (although somewhat technical and artificial) that shows how things can go wrong without it. Let the state space $\mathcal{S}$ be the unit interval $[0, 1]$. Let $B = \{2^{-n} : n = 1, 2, \ldots\}$. Define the distribution π to have probability mass 1/2 on the point 1 and density 1/2 on the rest of the interval, $[0, 1)$. For each state $x \notin B$, take the next-state distribution $P(x, \cdot)$ to be π. For $x = 2^{-n} \in B$, define $P(2^{-n}, \cdot)$ to have mass $\frac{2^{n+1}-2}{2^{n+1}-1}$ on the point $2^{-(n+1)}$ and the remaining mass $1/(2^{n+1} - 1)$ on the point 1. Show that the state 1 is an accessible atom, and that π is a stationary distribution for the chain. But what happens if we start out the chain in the state 1/2?*

[[For your convenience, a bit of helpful algebra: $\prod_{n=1}^{m} \frac{2^{n+1}-2}{2^{n+1}-1} = \frac{1}{2-2^{-m}}$.]]

1.10.2 Warm up for Harris chains

The purpose of this section is to warm up for the next section on Harris chains. If you are already feeling warm, you might find all this a bit slow and repetitious, in which case you might try skipping to the next section and see how it goes. If that section seems mysterious to you, you can always come back here then.

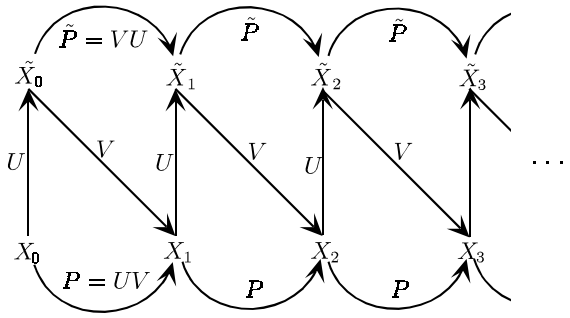
To illustrate the method of thinking we will see how the ideas work in some simple chains having finite state spaces. Of course, the ideas are not needed in order to obtain a Basic Limit Theorem for countable-state Markov chains; we have already done that! But we will use the ideas to extend the Basic Limit Theorem to more general state spaces.

(1.80) EXAMPLE. A lesson of exercise (1.5) [[***make this an example rather than an exercise?]] was that we can "lump" states if the transition probabilities out of those states are the same. That is, what characterizes a state x is really its next-state transition probabilities $P(x, \cdot)$, and if $P(x, \cdot) = P(y, \cdot)$, then we may combine the two states x and y into one state and still have a Markov chain. In a sense, if we have just made a transition and are told that the chain went to either x or y and we are wondering which, it really doesn't matter, in the sense that it makes no difference to our probabilistic predictions of the future path of the chain. In general, suppose there is a set R of states all having the same next-state transition probabilites; that is, suppose $P(x, \cdot) = P(y, \cdot)$ for all $x, y \in R$.

Then we may lump the states in R into a new state α, say. Whenever the X chain enters the set R, that is, whenever it occupies a state in the set R, we will say that the chain $\tilde{X}$ enters the state α. For example, given a chain $X_0, X_1, \ldots$ having transition matrix $P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} .1 & .5 & .4 \\ .3 & .1 & .6 \\ .3 & .1 & .6 \end{pmatrix} \end{array}$, states 2 and 3 may be lumped into one state α. That is, if we just keep track of visits to state 1 and state α, defining $\tilde{X}_t$ by

$$\tilde{X}_t = \begin{cases} 1 & \text{if } X_t = 1 \\ \alpha & \text{if } X_t \in \{2, 3\} \end{cases},$$

the process $\tilde{X}_0, \tilde{X}_1, \ldots$ is a Markov chain in its own right, with transition matrix $\tilde{P} = \begin{array}{c} \\ 1 \\ \alpha \end{array} \begin{array}{c} \begin{array}{cc} 1 & \alpha \end{array} \\ \begin{pmatrix} .1 & .9 \\ .3 & .7 \end{pmatrix} \end{array}$. In fact, we can combine the processes together to form the interlaced sequence $X_0, \tilde{X}_0, X_1, \tilde{X}_1, \ldots$, which is also a Markov chain, although it is time-inhomogeneous. The transitions from X_t to $\tilde{X}_t$ use the matrix $U = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cc} 1 & \alpha \end{array} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \end{array}$, and the transitions from $\tilde{X}_t$ to X_{t+1} use the matrix $V = \begin{array}{c} \\ 1 \\ \alpha \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} .1 & .5 & .4 \\ .3 & .1 & .6 \end{pmatrix} \end{array}$. Note that $UV = P$ and $VU = \tilde{P}$. □



(1.81) FIGURE. *A tricky but useful way of thinking of running the chain.*

This edifice we have erected on top of the given chain $X_0, X_1, \ldots$ is an unnecessarily complicated way of thinking about this particular chain, but this style of thinking will

be used for the general Basic Limit Theorem. This sort of lumping of states becomes particularly important in uncountably infinite state spaces, where each individual state may be hit with probability 0 while *sets* of states can be hit with positive probability. In such a case, by considering a set of states as a new lumped state, we can produce an atom.

Next let us look at a case where there is no pair of states with exactly the same transition probabilities. This is the typical case; for example, in Example ..., no two states have the same next-state transition probabilities. But nearby states have nearly the same transition probabilities. This will allow us to use a modified version of the trick above. We'll see that it is enough for a set of states to have a common "component."

(1.82) EXAMPLE. Consider the matrix $P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} .1 & .5 & .4 \\ .4 & 0 & .6 \\ .3 & .2 & .5 \end{pmatrix} \end{array}$, and suppose we are interested in lumping the states in the set $R = \{2, 3\}$. Now since $P(2, \cdot) \neq P(3, \cdot)$ things are not as simple as before. But note that rows 2 and 3 of P are both at least $(.3, 0, .5) = 0.8(.375, 0, .625)$. In fact,

$$P(2, \cdot) = (.4, 0, .6) = (.3, 0, .5) + (.1, 0, .1) = 0.8(.375, 0, .625) + 0.2(.5, 0, .5)$$

and

$$P(3, \cdot) = (3., .2, .5) = (.3, 0, .5) + (0, .2, 0) = 0.8(.375, 0, .625) + 0.2(0, 1, 0).$$

These equations express each of the distributions $P(2, \cdot)$ and $P(3, \cdot)$ as a mixture of the distribution $(.375, 0, .625)$ with some other distribution. In other words, both distributions $P(2, \cdot)$ and $P(3, \cdot)$ share the common "component" $0.8(.375, 0, .625)$. A useful interpretion of these equations is as follows. Suppose we have access to a biased coin having probability 0.8 of Heads and probability 0.2 of Tails. In order to generate the next state of the chain, given the present state is 2, we start by tossing the coin. If we get Heads, we then draw from the distribution $(.375, 0, .625)$, and if we get Tails, we draw from the distribution $(.5, 0, .5)$. Similarly, if we are now in state 3, we can generate the next state by tossing the same coin, drawing from the distribution $(.375, 0, .625)$ if we get Heads, and drawing from the distribution $(0, 1, 0)$ if we get Tails.

With this description, there are now two scenarios under which we use precisely the same distribution [[i.e., $(.375, 0, .625)$]] to generate the next state:

1. Enter state 2 and get Heads from coin toss
2. Enter state 3 and get Heads from coin toss

Since these two scenarios lead to the same next-state distribution, we can lump them together into a new state.

So here is another way to conceptualize the running of this chain. At time t, say the state is X_t. First we look to see whether we are in either of states 2 or 3, and if so we toss the biased coin, getting the outcome $C \in \{\text{Heads}, \text{Tails}\}$. Then define $\tilde{X}_t$ as follows:

$$\tilde{X}_t = \begin{cases} X_t & \text{if } X_t = 1, \text{ or if } X_t \in \{2, 3\} \text{ and } C = \text{Tails} \\ \alpha & \text{if } X_t \in \{2, 3\} \text{ and } C = \text{Heads} \end{cases}$$

We can use $\tilde{X}_t$ to generate the next state X_{t+1} as follows. If $\tilde{X}_t = 1$, we draw X_{t+1} from the probability mass function $(.1, .5, .4)$. If $\tilde{X}_t = 2$, then we know that X_t was 2 and C came out as Tails, so we use the mass function $(.5, 0, .5)$. Similarly, if $\tilde{X}_t = 3$, we use the mass function $(0, 1, 0)$. Finally, if $\tilde{X}_t = \alpha$, we know that X_t was either 2 or 3 and $C =$ Heads, so we use the mass function $(.375, 0, .625)$.

Again we have decomposed each transition of the given chain, according to P, into 2 stages, as depicted in Figure (1.81). These stages make transitions according to the matrices U and V, given by

$$U = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & \alpha \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & .2 & 0 & .8 \\ 0 & 0 & .2 & .8 \end{pmatrix} \end{array}, \quad V = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \alpha \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} .1 & .5 & .4 \\ .5 & 0 & .5 \\ 0 & 1 & 0 \\ .375 & 0 & .625 \end{pmatrix} \end{array}.$$

We started with a set of states $R = \{2, 3\}$. For each $i \in R$, we then wrote $P(i, \cdot)$ as a mixture of some fixed probability mass function $\rho = (.375, 0, .625)$ with some other probability mass function $Q(i, \cdot)$ [[in our example $Q(2, \cdot) = (.5, 0, .5)$ and $Q(3, \cdot) = (0, 1, 0)$]]. $P(i, \cdot) = 0.8\rho + 0.2Q(i, \cdot)$.

We have broken down each transition of the chain into two stages. Starting from the state X_t in the first stage we note whether or not X_t is in the set R, and if so we toss the biased coin. If the coin toss comes up Heads, we move to state α, and otherwise we stay where we are; the result is the state we have called $\tilde{X}_t$. Then we draw the next state X_{t+1} from the appropriate distribution. The point is that we have introduced a new state α that we can reach by hitting *any state in the set* R and then getting a Heads from the coin toss. This is the key in general state spaces: if we can take the set R to be large enough, the set R will have positive probability of being hit, even though each individual state in R may have probability 0 of being hit. And if R is hit with positive probability, then so is α, since hitting α only requires hitting R and a Heads from the coin toss.

Note also that we could have chosen R in different ways. For example, consider taking R to be the whole state space $\{1, 2, 3\}$. In that case we have

$$P(i, \cdot) \geq (.1, 0, .4) = .5(.2, 0, .8) \text{ for all } i \in R.$$

So we can take $\rho = (.2, 0, .8)$ and for each $i \in R = \{1, 2, 3\}$ write $P(i, \cdot)$ as a mixture

$$P(i, \cdot) = 0.5\rho + 0.5Q(i, \cdot),$$

where $Q(1, \cdot) = (0, 1, 0)$, $Q(2, \cdot) = (.6, 0, .4)$, and $Q(3, \cdot) = (.4, .4, .2)$. The way of running the chain that corresponds to this decomposition of the transition probabilities is as follows. Starting from any state X_t, toss a coin with $\mathbb{P}\{\text{Heads}\} = 0.5$. If Heads, define $\tilde{X}_t = \alpha$, with $\tilde{X}_t = X_t$ otherwise. Then choose X_{t+1} according to the probability mass function ρ if $\tilde{X}_t = \alpha$ and according to $Q(i, \cdot)$ if $\tilde{X}_t = i \in \mathcal{S}$. □

(1.83) EXERCISE.

(a) Suppose we have a finite-state Markov chain and we are considering taking our set R to consist of 2 states $R = \{i, j\}$. Express ρ and "$\mathbb{P}\{$Heads$\}$" in terms of the ith and jth rows of the probability transition matrix of the chain. In particular, show that $p = 1 - \|P(i,\cdot) - P(j,\cdot)\|$.

(b) Consider the frog chain. What happens when we try to take $R = \{1, 2\}$?

1.10.3 Harris Chains

A Markov chain $\{X_t\}$ with transition kernel P is a ***Harris chain*** if there is a set $R \subseteq \mathbb{S}$, a probability measure ρ on $\mathbb{S}$, and a positive number ϵ such that

(1) $\mathbb{P}_x\{X_t \in R \text{ for some } t \geq 0\} > 0$ for all $x \in \mathbb{S}$

(2) For all states $x \in R$ and all subsets $A \subseteq \mathbb{S}$, $P(x, A) \geq \epsilon\rho(A)$.

Conditions (1) and (2) pull in opposite directions: Roughly speaking, (1) wants the set R to be large, while (2) wants R to be small. Condition (1) requires that R be accessible from each state $x \in \mathbb{S}$. For example, (1) is satisfied trivially by taking R to be the whole state space $\mathbb{S}$, but in that case (2) becomes a very demanding condition, asking for $P(x,\cdot) \geq \epsilon\rho(\cdot)$ to hold for all states $x \in \mathbb{S}$. On the other hand, (2) is satisfied trivially if we take R to be any singleton $\{x_1\}$: just take $\rho(\cdot)$ to be $P(x_1,\cdot)$ (and take $\epsilon = 0.9$, for example). But in many examples each singleton is hit with probability 0, so that no singleton choice for R will satisfy condition (1). A Harris chain is one for which there is a set R that is simultaneously large enough to satisfy (1) but small enough to satisfy (2).

Let's think a bit about the interpretation of (2). What does this inequality tell us? Writing

$$P(x, A) = \epsilon\,[\rho(A)] + (1 - \epsilon)\left[\frac{P(x, A) - \epsilon\rho(A)}{1 - \epsilon}\right] =: \epsilon\rho(A) + (1 - \epsilon)Q(x, A),$$

we have expressed the distribution $P(x,\cdot)$ as a mixture of two probability distributions ρ and $Q(x,\cdot)$, where $Q(x,\cdot)$ is defined by $Q(x, A) = [P(x, A) - \epsilon\rho(A)]/(1 - \epsilon)$. Note that $Q(x,\cdot)$ is indeed a probability measure; for example, $Q(x, A) \geq 0$ by the assumption that $P(x, A) \geq \epsilon\rho(A)$, and $Q(x, \mathbb{S}) = 1$ because we have divided by the appropriate quantity $(1 - \epsilon)$ in the defining $Q(x,\cdot)$. Thus, we can simulate a draw from the distribution $P(x,\cdot)$ by the following procedure.

- Flip a "coin" having $\mathbb{P}(\text{heads}) = \epsilon$ and $\mathbb{P}(\text{tails}) = 1 - \epsilon$.
- If the outcome is heads, take a random draw from the distribution ρ.
- If the outcome is tails, take a draw from the distribution $Q(x,\cdot)$.

It is useful to imagine essentially the same process in another slightly different way, on a slightly different state space. Let us adjoin an additional state, α, to the given state space $\mathcal{S}$, obtaining the new state space $\tilde{\mathcal{S}} = \mathcal{S} \cup \{\alpha\}$. This new state α will be our accessible atom. We will say that the new chain visits the state α whenever the old chain enters the set R and the coin flip turns up heads. Thus, after the state α is entered, we know that the next state will be distributed according to the distribution ρ; note that this distribution is the same for all $x \in R$. When the chain enters the state $x \in R$ and the coin flip turns up tails, the next state is chosen according to the distribution $Q(x, \cdot)$.

To put all of this together, consider a Markov chain $X_0, \tilde{X}_0, X_1, \tilde{X}_1, \ldots$ generated recursively as follows. Suppose we are at time t, and we have already generated the value of X_t, and we are about to generate $\tilde{X}_t$. If $X_t \in R^c = \mathcal{S} - R$, then $\tilde{X}_t = X_t$. If $X_t \in R$, then we toss a coin. If the toss comes up heads, which happens with probability ϵ, then $\tilde{X}_t = \alpha$. If the toss comes up tails, then $\tilde{X}_t = X_t$. Next we use the value of $\tilde{X}_t$ to generate X_{t+1}. If $\tilde{X}_t = \alpha$ then X_{t+1} is chosen from the distribution ρ. If $\tilde{X}_t \in R$ then X_{t+1} is chosen from the distribution $Q(X_t, \cdot)$. If $\tilde{X}_t \neq \alpha$ and $\tilde{X}_t \notin R$ then X_{t+1} is chosen from the distribution $P(X_t, \cdot)$.

In other words, again we have imbedded our given Markov chain in the structure shown in Figure (1.81), with the transition kernels U and V given by

$$\begin{aligned}
\text{For } x \in R\text{:}\quad & U(x, \{\alpha\}) = \epsilon, U(x, \{x\}) = 1 - \epsilon \\
\text{For } x \in \mathcal{S} - R\text{:}\quad & U(x, \{x\}) = 1 \\
& V(\alpha, A) = \rho(A) \\
\text{For } x \in R\text{:}\quad & V(x, A) = Q(x, A) \\
\text{For } x \in \mathcal{S} - R\text{:}\quad & V(x, A) = P(x, A).
\end{aligned}$$

The sequence $X_0, \tilde{X}_0, X_1, \tilde{X}_1, \ldots$ is a time-inhomogeneous Markov chain; the transition kernel U used in going from X_t to $\tilde{X}_t$ is different from the kernel V used in going from $\tilde{X}_t$ to X_{t+1}. Note that $X_t \in \mathcal{S}$ and $\tilde{X}_t \in \tilde{\mathcal{S}}$ for all t. The sequence $X_0, X_1, \ldots$ is a time-homogeneous Markov chain on $\mathcal{S}$ with transition kernel UV, defined by

$$(UV)(x, B) = \int U(x, dy) V(y, B).$$

We claim that $UV = P$. If $x \in \mathcal{S} - R$ then $U(x, \cdot)$ is point mass on x, so that $(UV)(x, B) = V(x, B) = P(x, B)$. If $x \in R$ then $U(x, \cdot)$ puts probability ϵ on the point α and probability $1 - \epsilon$ on the point x, so that

$$\begin{aligned}
(UV)(x, B) &= \epsilon V(\alpha, B) + (1 - \epsilon) V(x, B) \\
&= \epsilon \rho(B) + (1 - \epsilon) Q(x, B) = P(x, B).
\end{aligned}$$

The sequence $\tilde{X}_0, \tilde{X}_1, \ldots$ is a time-homogeneous Markov chain, with transition kernel $VU =: \tilde{P}$.

(1.84) EXERCISE. *Write down the transition kernel $\tilde{P}$ in terms of the information given in the problem.*

If X_t has distribution π_t on $\mathbb{S}$, then $\tilde{X}_t$ has distribution $\tilde{\pi}_t = \pi_t U$ on $\widetilde{\mathbb{S}}$.

Finally, here is our Basic Limit Theorem for Harris chains. As usual, the statement involves an aperiodicity condition. Letting $G = \{t \geq 1 : \mathbb{P}_\rho\{X_{t-1} \in R\} > 0\}$, we say the chain is aperiodic if $\gcd(G) = 1$. For example, as a simple sufficient condition, if $\rho(R) > 0$, then the set G contains 1, so that the chain is aperiodic.

(1.85) THEOREM. *Let $\{X_t\}$ be an aperiodic Harris chain having a stationary distribution π. Let π_t denote the distribution of X_t and let the initial distribution π_0 be absolutely continuous with respect to π. Then $\|\pi_t - \pi\| \to 0$ as $t \to \infty$.*

PROOF: We are given the Harris chain $\{X_t\}$ with transition kernel P. Suppose we are also given a set R, probability measure ρ, and number $\epsilon \in (0,1)$ as in the definition of a Harris chain. As discussed above, these determine transition kernels U and V with $P = UV$ and $\tilde{P} = VU$, and we will study the chain $X_0, \tilde{X}_0, X_1, \tilde{X}_1, \ldots$. We are assuming that $\{X_t\}$ has a stationary distribution π, and we now know that $\{\tilde{X}_t\}$ has corresponding stationary distribution $\tilde{\pi} = \pi U$. By the definition of the Harris chain $\{X_t\}$, the state α is an accessible atom for $\{\tilde{X}_t\}$, and the aperiodicity assumption implies that α is aperiodic. ⟦***WHY? EXPLAIN THIS.⟧ Defining $\tilde{\pi}_0 = \pi_0 U$, we see that $\tilde{\pi}_0$ is absolutely continuous with respect to $\tilde{\pi}$. Therefore, by Theorem (1.78) we have $\|\tilde{\pi}_t - \tilde{\pi}\| \to 0$, where $\tilde{\pi}_t$ denotes the distribution of $\tilde{X}_t$. But

$$\tilde{\pi}V = (\pi U)V = \pi(UV) = \pi P = \pi.$$

Thus, since

$$\|\pi_{t+1} - \pi\| = \|\tilde{\pi}_t V - \tilde{\pi}V\| \leq \|\tilde{\pi}_t - \tilde{\pi}\|,$$

we have $\|\pi_{t+1} - \pi\| \to 0$ as $t \to \infty$. □

*** NOTE: Argue somewhere that $\|\lambda P - \mu P\| \leq \|\lambda - \mu\|$. Can use coupling. Consider chains $\{X_t\}$, $\{Y_t\}$ having transition rule P, with $X_0 \sim \lambda$ and $Y_0 \sim \mu$. Look at $\mathbb{P}\{X_1 = Y_1\}$, conditioning on whether or not $X_0 = Y_0$.

*** ALSO apply this stuff back to a Gibbs sampling example.

1.10.4 More about stationary distributions

*** Omit or incorporate in earlier sections?

Suppose the chain has a positive recurrent atom α, so that $\mathbb{E}_\alpha(T_\alpha) < \infty$. Define

$$\pi(A) = \frac{\mathbb{E}_\alpha\left[\sum_{t=0}^{T_\alpha - 1} I\{X_t \in A\}\right]}{\mathbb{E}_\alpha(T_\alpha)}. \tag{1.86}$$

What is this? Remember the I denotes an indicator random variable. The sum $\sum_{t=0}^{T_\alpha - 1} I\{X_t \in A\}$ is accumulating 0's and 1's as t ranges over the values $0, 1, \ldots, T_\alpha - 1$. So the sum is simply a count of the number of times that $X_t \in A$ holds for t between 0 and $T_\alpha - 1$. In other words, the sum is the number of visits made by $X_0, \ldots, X_{T_\alpha - 1}$ to the set A, and the numerator of $\pi(A)$ is the expected number of such visits. Think again of the "cycle" idea, where a cycle is now a portion of the Markov chain path between successive

visits to the state α. Then $\pi(A)$ is the expected number of times the chain visits the set A during a cycle, divided by the expected length of a cycle.

Now, T_a is a random variable, so the sum in (1.86) is running over a random number of terms. That looks a bit hard to work with, but we can use the following standard and useful trick, which should be your first reaction when you see sums like this: we make the summation sign run over all possible t values and introduce another indicator function to restrict the sum to the values of t that we want. That is,

$$\sum_{t=0}^{T_\alpha - 1} I\{X_t \in A\} = \sum_{t=0}^{\infty} I\{X_t \in A\} I\{t < T_a\} = \sum_{t=0}^{\infty} I\{X_t \in A, T_a > t\}.$$

Taking the expected value, since the expected value of an indicator random variable is its probability, we can write π in the equivalent form

$$\pi(A) = \frac{\sum_{t=0}^{\infty} \mathbb{P}_\alpha\{X_t \in A, T_\alpha > t\}}{\mathbb{E}_\alpha(T_\alpha)}. \tag{1.87}$$

The manipulation from (1.86) to (1.87) is so fundamental and often used in probability that you will often see it used without any comment. It is a trick that is well worth mastering and remembering.

(1.88) PROPOSITION. *Let $\{X_t\}$ be a Markov chain with a positive recurrent atom α, and define*

$$\pi(A) = \frac{\mathbb{E}_\alpha\left[\sum_{t=0}^{T_\alpha - 1} I\{X_t \in A\}\right]}{\mathbb{E}_\alpha(T_\alpha)} = \frac{\sum_{t=0}^{\infty} \mathbb{P}_\alpha\{X_t \in A, T_\alpha > t\}}{\mathbb{E}_\alpha(T_\alpha)}.$$

Then π is a stationary distribution for $\{X_t\}$.

PROOF: Clearly π is a probability distribution. We want to show that $\int P(x, A)\pi(dx) = \pi(A)$. Defining $\mu(A) = \mathbb{E}_\alpha(T_\alpha)\pi(A)$, we want to show that $\int P(x, A)\mu(dx) = \mu(A)$. We have

$$\int P(x, A)\mu(dx) = \sum_{t=0}^{\infty} \int \mathbb{P}_\alpha\{X_t \in dx, T_\alpha > t\} P(x, A).$$

But

$$P(x, A) = \mathbb{P}_\alpha\{X_{t+1} \in A \mid X_t = x\} = \mathbb{P}_\alpha\{X_{t+1} \in A \mid X_t = x, T_\alpha > t\},$$

where the last equality holds by the Markov property, because the event $\{T_\alpha > t\} = \{T_\alpha \leq t\}^c$ depends only on the random variables $X_0, \ldots, X_t$. (That is, given the precise information about the state $X_t = x$, we can throw away the information $T_\alpha > t$.) So

$$\begin{aligned} \int P(x, A)\mu(dx) &= \sum_{t=0}^{\infty} \int \mathbb{P}_\alpha\{X_t \in dx, T_\alpha > t\} \mathbb{P}_\alpha\{X_{t+1} \in A \mid X_t = x, T_\alpha > t\} \\ &= \sum_{t=0}^{\infty} \mathbb{P}_\alpha\{X_{t+1} \in A, T_\alpha > t\} \end{aligned}$$

$$= \mathbb{E}_\alpha \left[\sum_{t=0}^{T_\alpha - 1} I\{X_{t+1} \in A\} \right]$$
$$= \mathbb{E}_\alpha \left[\sum_{t=1}^{T_\alpha} I\{X_t \in A\} \right].$$

That is, $\int P(x, A)\mu(dx)$ is the expected number of visits made by the segment $(X_1, \ldots, X_{T_\alpha})$ to the set A. Is this the same as $\mu(A)$, which is the expected number of visits made by the segment $(X_0, \ldots, X_{T_\alpha - 1})$ to the set A? The answer is yes! In fact, since $X_0 = X_{T_\alpha} = \alpha$, the two segments $(X_1, \ldots, X_{T_\alpha}) = (X_1, \ldots, X_{T_\alpha - 1}, \alpha)$ and $(X_0, \ldots, X_{T_\alpha - 1}) = (\alpha, X_1, \ldots, X_{T_\alpha - 1})$ consist of precisely the same states, just visited in a different order. Of course the mere difference in ordering leaves the number of visits to the set A unchanged between the two segments. □

(1.89) PROPOSITION. *Suppose a Markov chain has an accessible atom α and a stationary distribution π. Then $\pi\{\alpha\} = 1/\mathbb{E}_\alpha(T_\alpha)$.*

PROOF: By the same proof as the SLLN before, using the cycle idea, we know that if the chain is started in the state α, then $(1/n)\sum_{t=1}^n I\{X_t = \alpha\} \to 1/\mathbb{E}_\alpha(T_\alpha)$ with probability 1. Combining this with Proposition (1.76), here is what we know. If the chain is started out in the distribution π, then with probability 1 it hits α at some finite time, after which, with probability 1, the long run fraction of visits to α converges to $1/\mathbb{E}_\alpha(T_\alpha)$. We have used this type of reasoning before: the finite amount of time it takes the chain to hit α does not have any effect on the limiting long-run fraction of time the chain spends in the state α. Thus, for a chain started in the distribution π,

$$\mathbb{P}_\pi \left\{ \lim_{n\to\infty} (1/n) \sum_{t=1}^n I\{X_t = \alpha\} = 1/\mathbb{E}_\alpha(T_\alpha) \right\} = 1.$$

By the Bounded Convergence Theorem,

$$\mathbb{E}_\pi \left\{ (1/n) \sum_{t=1}^n I\{X_t = \alpha\} \right\} \to 1/\mathbb{E}_\alpha(T_\alpha)$$

as $n \to \infty$. But for each n,

$$\mathbb{E}_\pi \left\{ (1/n) \sum_{t=1}^n I\{X_t = \alpha\} \right\} = (1/n) \sum_{t=1}^n \mathbb{P}_\pi\{X_t = \alpha\} = \pi\{\alpha\}.$$

This, $\pi\{\alpha\} = 1/\mathbb{E}_\alpha(T_\alpha)$. □

*** ALTERNATIVELY, do it this way.....

(1.90) THEOREM. *Suppose the chain $\{X_t\}$ has an accessible atom α and a stationary distribution π. Then*

1. $\pi\{\alpha\} > 0$
2. *α is positive recurrent:* $\mathbb{E}_\alpha(T_\alpha) < \infty$
3. *For all* $A \in \mathcal{A}$,
$$\pi(A) = \frac{1}{\mathbb{E}_\alpha(T_\alpha)} \mathbb{E}_\alpha \sum_{t=0}^{T_\alpha - 1} I\{X_t \in A\}$$
4. *For π-a.a. x,* $\mathbb{P}_x\{T_\alpha < \infty\} = 1$.

PROOF:

1. Since α is accessible, for all states x, $P^t(x,\alpha) > 0$ for some t. That is, defining $G_t = \{x : P^t(x,\alpha) > 0\}$, we have $\bigcup G_t = \mathbb{S}$. So there is an n such that $\pi(G_n) > 0$, which gives
$$\pi\{\alpha\} = \int \pi(dx) P^n(x,\alpha) \geq \int_{G_n} \pi(dx) P^n(x,\alpha).$$
The last expression is an integral of a positive function over a set of positive measure, so it is positive.

2. Let $A \in \mathcal{A}$ be arbitrary. Start with the general decomposition
$$\begin{aligned}\mathbb{P}\{X_n \in A\} = &\mathbb{P}\{X_{n-1} = \alpha, X_n \in A\} + \mathbb{P}\{X_{n-2} = \alpha, X_{n-1} \neq \alpha, X_n \in A\} \\ &+ \cdots + \mathbb{P}\{X_0 = \alpha, X_1 \neq \alpha, \ldots, X_{n-1} \neq \alpha, X_n \in A\} \\ &+ \mathbb{P}\{X_0 \neq \alpha, X_1 \neq \alpha, \ldots, X_{n-1} \neq \alpha, X_n \in A\}.\end{aligned}$$
For a stationary chain this becomes
$$\begin{aligned}\pi(A) &= \pi\{\alpha\}\mathbb{P}_\alpha\{X_1 \in A\} + \pi\{\alpha\}\mathbb{P}_\alpha\{X_1 \neq \alpha, X_2 \in A\} \\ &\quad + \cdots + \pi\{\alpha\}\mathbb{P}_\alpha\{X_1 \neq \alpha, \ldots, X_{n-1} \neq \alpha, X_n \in A\} \\ &\quad + \mathbb{P}_\pi\{X_0 \neq \alpha, X_1 \neq \alpha, \ldots, X_{n-1} \neq \alpha, X_n \in A\} \\ &= \pi\{\alpha\}\mathbb{P}_\alpha\{X_1 \in A\} + \pi\{\alpha\}\mathbb{P}_\alpha\{X_2 \in A, T_\alpha \geq 2\} \\ &\quad + \cdots + \pi\{\alpha\}\mathbb{P}_\alpha\{X_n \in A, T_\alpha \geq n\} + \mathbb{P}_\pi\{X_0 \neq \alpha, \ldots, X_{n-1} \neq \alpha, X_n \in A\}.\end{aligned}$$

Dropping the last term, we get the inequality
$$\pi(A) \geq \pi\{\alpha\} \sum_{t=1}^{n} \mathbb{P}_\alpha\{X_t \in A, T_\alpha \geq t\} = \frac{1}{\mathbb{E}_\alpha(T_\alpha)} \sum_{t=1}^{n} \mathbb{P}_\alpha\{X_t \in A, T_\alpha \geq t\},$$
and since this holds for all n, we have
$$\pi(A) \geq \frac{1}{\mathbb{E}_\alpha(T_\alpha)} \sum_{t=1}^{\infty} \mathbb{P}_\alpha\{X_t \in A, T_\alpha \geq t\}. \tag{1.91}$$
Applying this last inequality to the choice $A = \mathbb{S}$, the whole state space, gives $1 \geq \pi\{\alpha\} \sum_{t=1}^{\infty} \mathbb{P}_\alpha\{T_\alpha \geq t\} = \pi\{\alpha\}\mathbb{E}_\alpha(T_\alpha)$, so that, since we know $\pi\{\alpha\}$ is strictly positive, $\mathbb{E}_\alpha(T_\alpha) \leq 1/\pi\{\alpha\} < \infty$.

3. Defining $\tilde{\pi}(A)$ to be the right-hand side of (1.91), we have $\pi(A) \geq \tilde{\pi}(A)$ for all A. So since both π and $\tilde{\pi}$ are probability distributions, we must have $\pi(A) = \tilde{\pi}(A)$ for all A. [[Why?]]

□

1.11 More notes to myself

1. Streamline general state space stuff.
2. Make notation consistent; e.g. is MC time index t or n? Probably should make it t throughout.
3. Include a Gibbs sampling example.
4. More detail on counting contingency tables; describe an actual simulation run.